A Comparison of Classification Techniques for Heart Attack Risk Prediction

Saumya Shah, Anas Tagui

Politechnika Warszawska

EARIN - Introduction to AI

10th June 2025

Abstract

This report presents a comparative analysis of K-Nearest Neighbours (KNN), Mixed Naive Bayes, and Neural Network classification techniques for predicting heart attack risk, based on a comprehensive dataset. Following robust preprocessing, including outlier removal and data splitting, each model's performance was rigorously evaluated using metrics such as F1-score, AUC, recall, precision, inference time, training time, generalisability, and interpretability. The Neural Network model demonstrated superior performance, achieving an F1-score of 0.9630 and an AUC of 0.97 on the test set, critically achieving perfect recall (1.0) with minimal false positives, making it the most suitable for clinical deployment due to its high accuracy, safety, and rapid inference. Mixed Naive Bayes also exhibited perfect recall and strong interpretability, positioning it as a viable, safety-focused alternative, despite a higher false positive rate. Conversely, the KNN classifier proved unsuitable for medical application due to unacceptable false negatives and slow inference times. This report underscores the importance of judicious model selection for sensitive medical diagnostic tasks.

Keywords: Heart Attack Risk Prediction, Classification Techniques, K-Nearest Neighbours, Naive Bayes, Neural Networks

Models

Overview

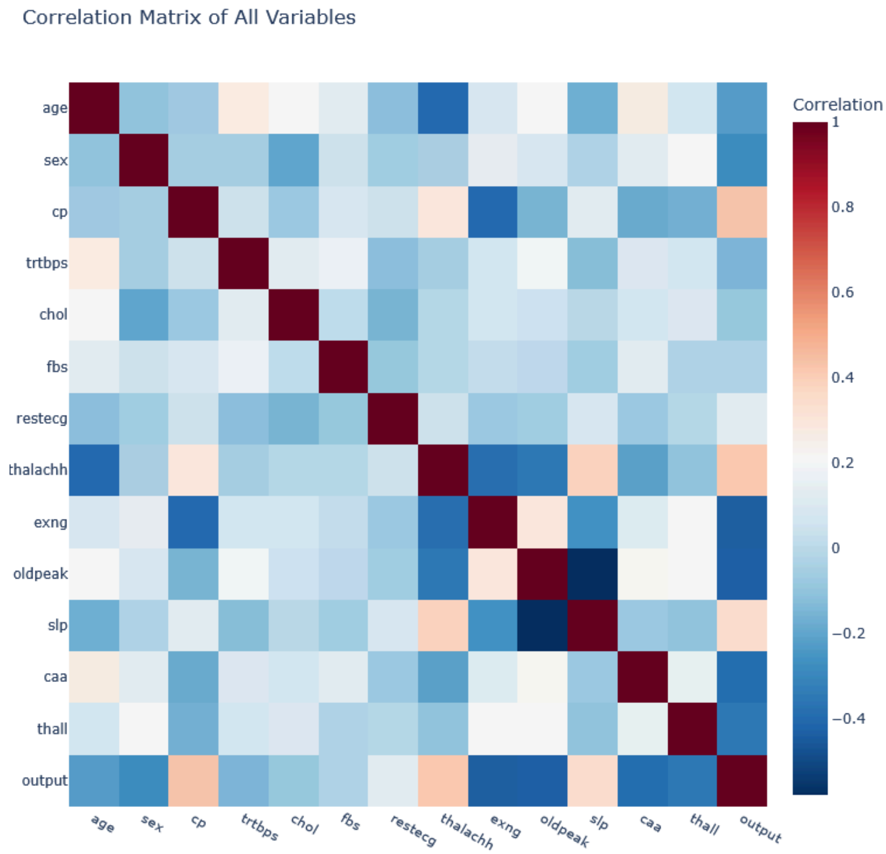
General Preprocessing Results

The initial data assessment reveals no missing or null values across any columns, eliminating the need for imputation or cleaning up of those values. While there were ordinal and nominal features, they were already label-encoded, eliminating the need for further processing.

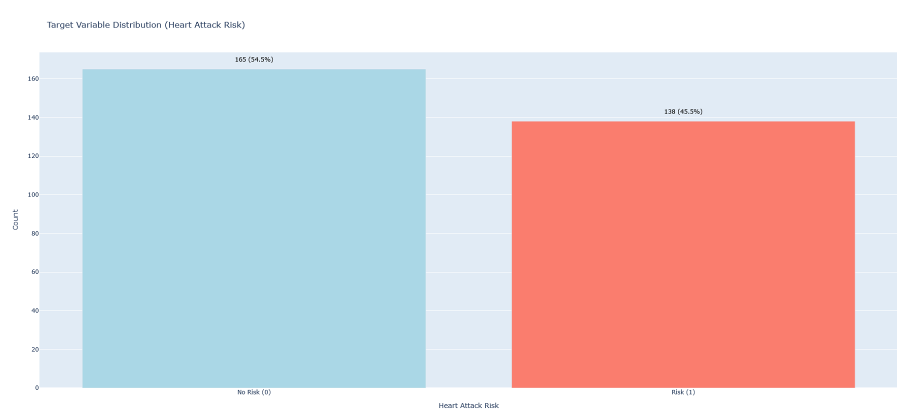
To ensure the robustness of the model and prevent extreme values from disproportionately influencing results, outlier detection and removal were performed. The Interquartile Range (IQR) was calculated for all numerical columns, and 19 data points falling outside the established lower and upper bounds were removed.

```
Outlier Analysis (IQR method):
==================================================
age: 0 outliers (0.0%)
  Bounds: [27.25, 81.25]
trtbps: 9 outliers (3.0%)
  Bounds: [90.00, 170.00]
chol: 5 outliers (1.7%)
  Bounds: [115.75, 369.75]
thalachh: 1 outliers (0.3%)
  Bounds: [84.75, 214.75]
oldpeak: 5 outliers (1.7%)
  Bounds: [-2.40, 4.00]
```

Given the relatively moderate dimensionality of the dataset (13 features), dimensionality reduction techniques were explored. However, no strong correlations were found between features, so all features were retained.



The imbalance detected between ‘risk’ and ‘no risk’ classes was insignificant, so class balancing techniques were required.



The dataset was split into three sets: 80% of the data was the training set, 10% was the validation set, and the remaining was the testing set.

K-Nearest Neighbours Classifier

Description

The K-Nearest Neighbours algorithm is a non-parametric instance-based classifier that operates without an explicit training phase; instead, it "memorises" the entire training dataset. To classify a new data point, KNN calculates its distance to every point in the training set, identifies the 'K' closest data points, and assigns the class label that represents the majority among these 'K' neighbours, effectively making a classification decision based on local proximity within the feature space.

Preprocessing

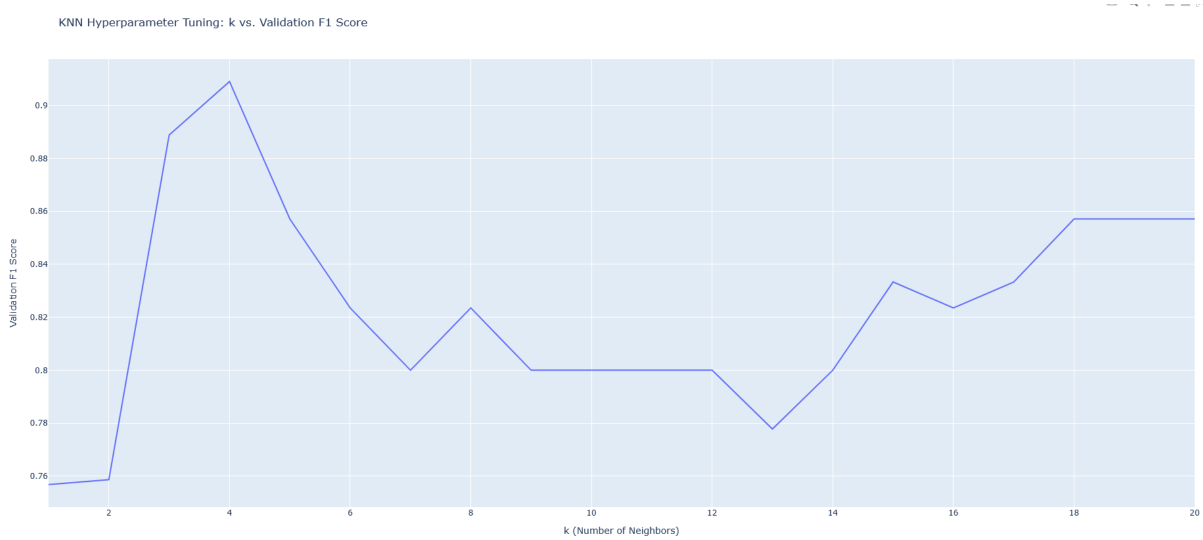
Given its reliance on distance computations, feature scaling is an important preprocessing step that ensures all features contribute equally to distance calculations by bringing them to a comparable range. However, because of our use of Gower distance, which uses range normalisation, scaling was not performed. Instead, the range was calculated and used during distance computations.

Furthermore, outlier removal is also required to prevent extreme values from distorting distance metrics and influencing neighbour selection. Dimensionality Reduction techniques (like PCA) are also required to mitigate the "curse of dimensionality," which can severely degrade KNN's performance and efficiency by making all data points appear sparsely distributed. Both of these were explored and performed in the general preprocessing step, where possible.

Results

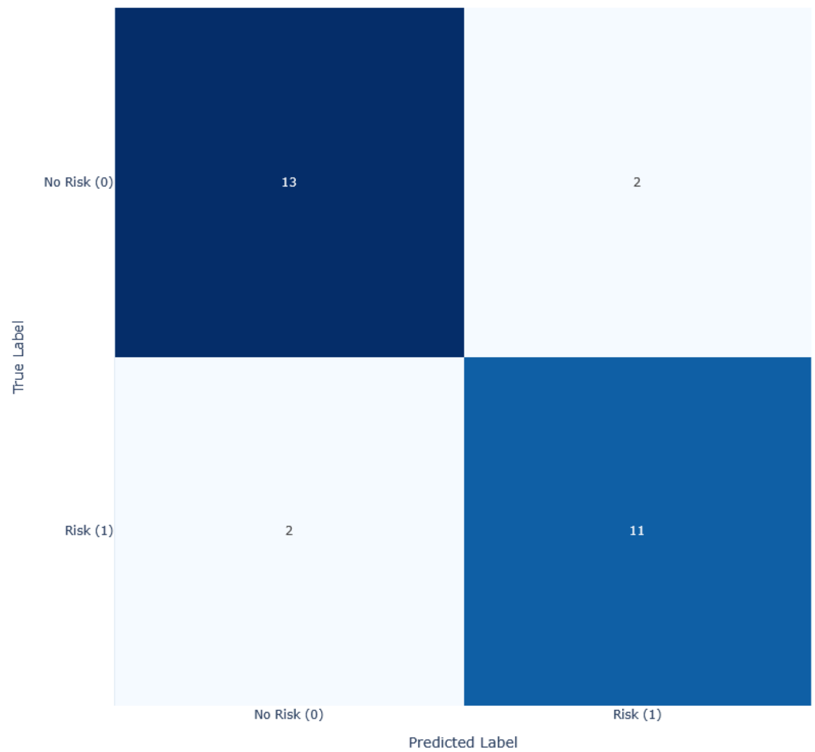
KNN with Gower distance was used due to the different kinds of data in the dataset (nominal, ordinal, continuous). For the distance between records A and B, for each continuous or ordinal feature, their range-normalised absolute distance was considered. For each nominal feature, the distance was considered 1 if their categories were different, and 0 otherwise. This total distance estimate was then averaged across all features.

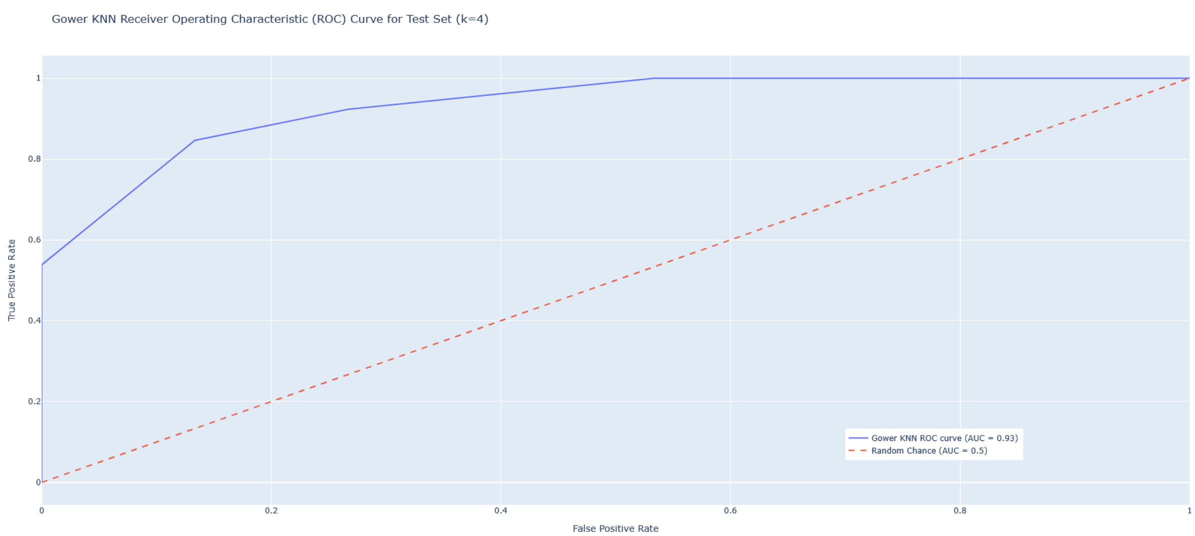
Then, automatic hyperparameter tuning was performed by fitting KNN classifiers for k-values between 1 and 20, and then calculating their F1 score on the validation set.



It was found that k = 4 was optimal, with a validation F1 score of 0.9091. The classifier corresponding to this k-value was then evaluated against the test set, which gave us our final F1 score of 0.8462.

```
 ## Summary for tuned KNN model ##
Final KNN Model - Validation F1 Score (k=4): 0.9091
Final KNN Model - Test F1 Score (k=4): 0.8462
Validation prediction:   0.04008410 seconds
Test prediction:         0.03674410 seconds
Samples per second (val): 699 samples/sec
Samples per second (test): 762 samples/sec
Samples per second (total): 729 samples/sec
Note: KNN does not have a training phase
```





The KNN model with Gower distance is a competent classifier, but falls short in the specific context of medical risk prediction.

An AUC of 0.91 and an F1-score of 0.8462 show that the model has solid predictive power. Its precision and recall are balanced, which shows a balance in being able to avoid false alarms while minimising missed risks.

However, it has unacceptable safety gaps; the model's relatively lower recall means it produced two false negatives. In a real-world scenario, this translates to failing to identify two patients who were genuinely at risk of a heart attack, which is a critical failure. The model is also considerably slower at making predictions than the other models, with an inference rate multiple orders of magnitude slower than the other two models. There is also a significant drop in performance between the validation and test sets, indicating that the model may not be very generalisable.

K-Nearest Neighbours Classifier

Description

The Naive Bayes classifier is a probabilistic classifier operating under the naive assumption that all features are conditionally independent of each other given the class label, and the data are independently distributed. For each class in our data, Naive Bayes models the distribution of each continuous feature as a distribution (either multinomial for ordinal/nominal data, or Gaussian for continuous data), characterised by its mean and variance within that class. When classifying a new data point, the algorithm calculates the likelihood of observing its feature values given each class, using the corresponding probability density functions. These likelihoods are then combined with the prior probabilities of each class, and the data point is assigned to the class with the highest resulting posterior probability.

Preprocessing

Outliers can significantly impact the estimation of means and variances for the classifier, potentially leading to inaccurate probability calculations. Therefore, outliers need to be identified and removed. In addition, highly correlated features are double-counted in the model, leading to overestimation of the importance of those features. Therefore, correlated features must be identified and removed. These two steps were performed in the general preprocessing step.

Results

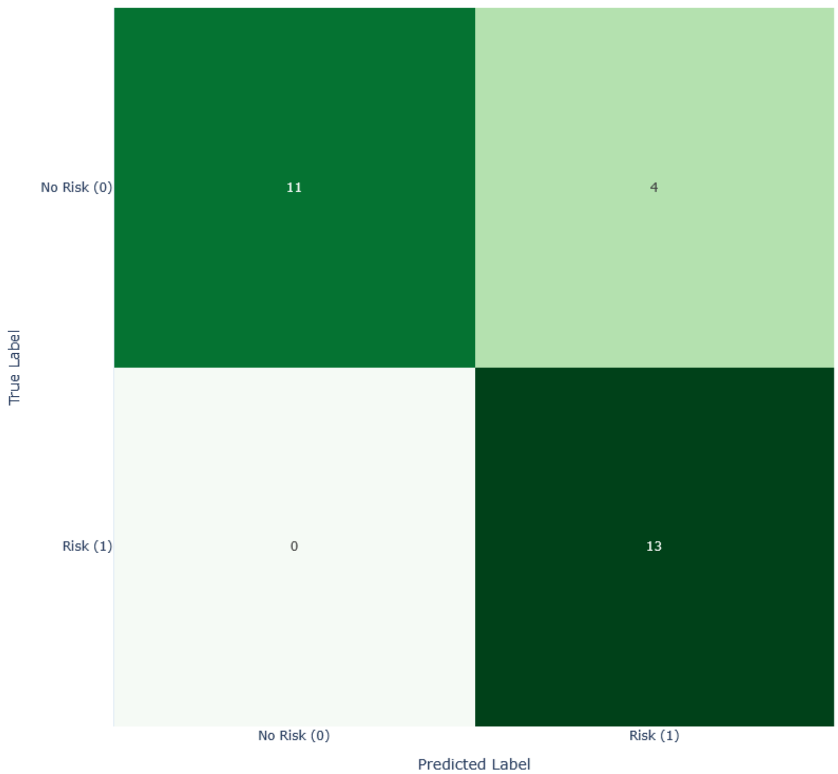
Mixed Naive Bayes (MNB) classification was performed by preprocessing and splitting the data, creating multinomial distributions for nominal and ordinal data, and Gaussian distributions for continuous data.

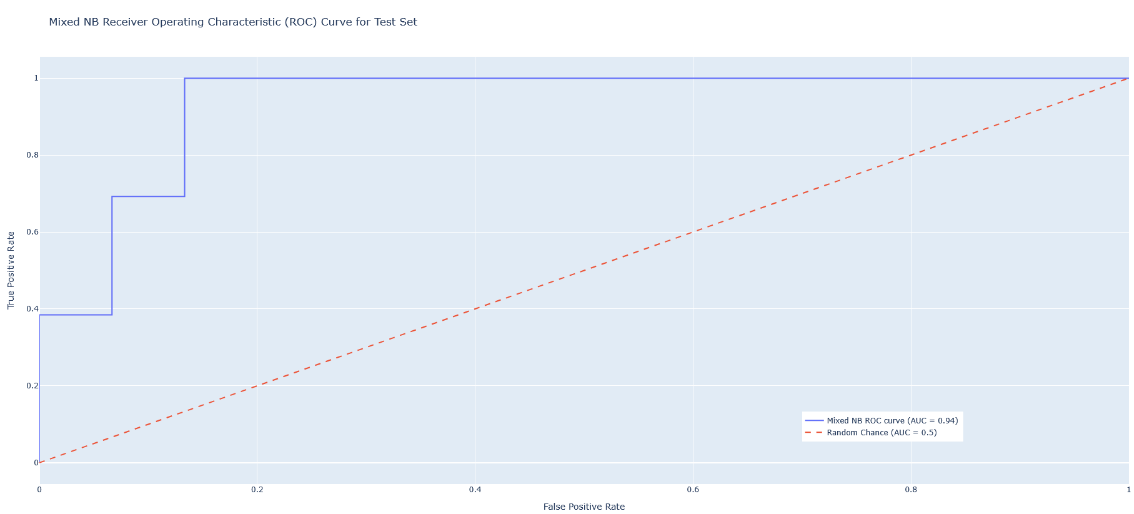
Then, automatic hyperparameter tuning was performed using grid search by fitting MNB classifiers for 20 var_smoothing and 20 alpha values between [10^{-12}, 10^{-6}] (for var_smoothing) and [10^{-3}, 10^{2}] (alpha), and then calculating their F1 score on the validation set.

It was found that var_smoothing = 10^{-12}, alpha = 10^{-3} was optimal, with a validation F1 score of 0.8947. The classifier corresponding to this k-value was then evaluated against the test set, which gave us our final F1 score of 0.8667.

```
Best parameters found:
  var_smoothing (Gaussian): 1.00e-12
  alpha (Multinomial): 1.00e-03
  Validation F1 Score: 0.8947
```

```
 ## Summary for Mixed Naive Bayes model ##
Final Mixed NB Model - Validation F1 Score: 0.8947
Final Mixed NB Model - Test F1 Score: 0.8667
Training time:            0.00182130 seconds
Validation prediction:    0.00082220 seconds
Test prediction:          0.00082390 seconds
Samples per second (val): 34055 samples/sec
Samples per second (test): 33985 samples/sec
Samples per second (total): 34020 samples/sec
```





The Naive Bayes model is an excellent classifier in practice even though its assumptions of conditional independence between features may be violated in theory.

It has a perfect safety profile for the context with a recall of 1.0, ensuring no at-risk patient was missed. It showed a negligible drop in performance between validation and testing, indicating it is very stable and generalisable. The model is fast to train and run, with the least time to first inference (training time + inference time) being the least of the three models.

With a small dataset, complex models (those with few assumptions and high flexibility) are highly prone to overfitting by memorising the training data, including its noise, meaning they have high variance. However, due to its strong biases and assumptions embedded in the form of the conditional independence assumption and the prior probability distributions, Naive Bayes trades some potential bias for a significant reduction in variance. This regularising effect helps increase generalisability when working with smaller datasets.

The model's primary weakness is a low precision of 0.765. This resulted in four false positives, meaning it incorrectly flagged four healthy patients as being at risk. While this is safer than missing a case, it is less efficient than other classifiers. Another challenge is the difficulty in selecting an appropriate distribution for each feature; while we were able to plot and observe an approximately Gaussian distribution for most continuous features in this dataset, this assumption may not hold for other features, and may lead to less accurate models.

Neural Network Classifier

Description

Neural Networks (NNs), especially deep learning architectures, are non-linear models composed of interconnected layers of neurons that learn complex patterns. For binary classification, an NN typically consists of an input layer (receiving the features), one or more hidden layers (where non-linear transformations occur via activation functions like ReLU), and an output layer. In binary classification, the output layer has a single neuron with a sigmoid activation function, which squashes its output to a value between 0 and 1, directly interpretable as the probability of belonging to the positive class. The network learns by iteratively adjusting its internal weights and biases through gradient descent to minimise a specified loss function.

Preprocessing

Feature scaling is essential to help stabilise and accelerate the convergence of the optimisation algorithm, preventing larger feature values from dominating the learning process. Dimensionality reduction techniques may also be required for very high-dimensional datasets to reduce noise, prevent overfitting, and potentially speed up training, particularly when data is sparse. Standard scaling was performed for this model, while dimensionality reduction was attempted in the general preprocessing step.

Results

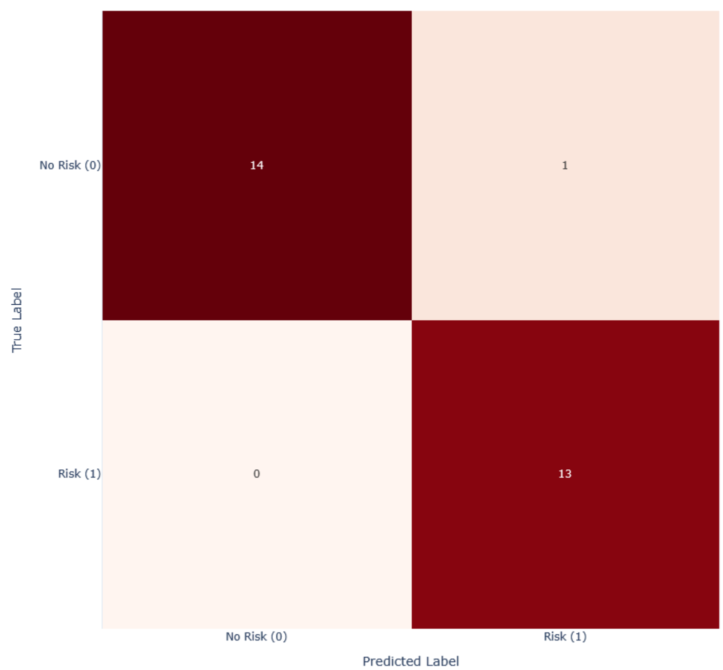
Neural network (NN) classification was performed by preprocessing and splitting the data, creating the network, and using backpropagation with BCE loss for training and ADAM for optimisation.

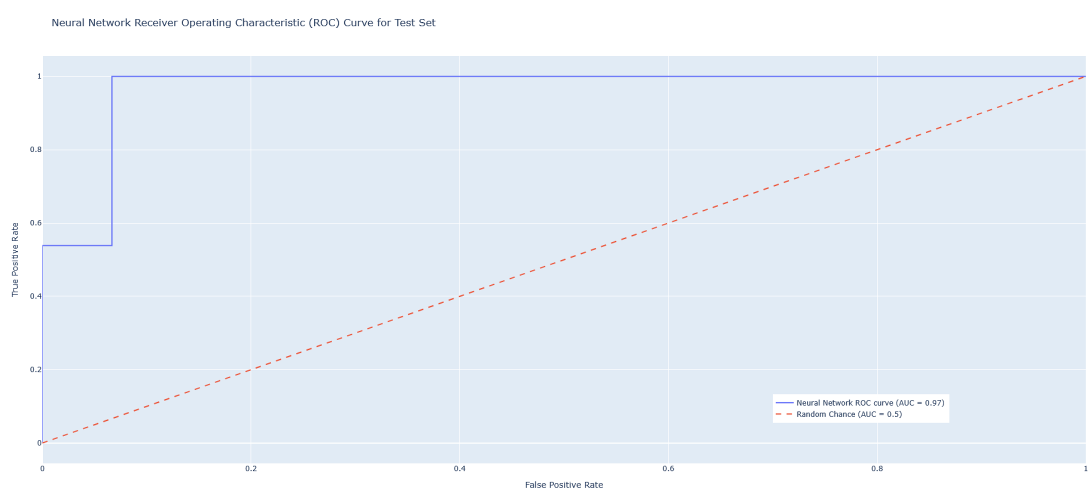
Then, automatic hyperparameter tuning was performed using grid search by fitting NN classifiers for 4 values of hidden layers (0, 1, 2, 3), 4 values for number of neurons in hidden layers (16, 32, 64, 128), and 2 values for learning rate (0.001 and 0.01).

```
Best parameters found:
  Hidden layers: 2
  Hidden neurons: 32
  Learning rate: 0.010
  Validation F1 Score: 0.9189
```

It was found that a classifier with 2 hidden layers, each with 32 neurons, trained with the initial learning rate of 0.01 was optimal, with a validation F1 score of 0.8947. The classifier corresponding to this k-value was then evaluated against the test set, which gave us our final F1 score of 0.8667.

```
 ## Summary for Neural Network model ##
Final Neural Network Model - Validation F1 Score: 0.9189
Final Neural Network Model - Test F1 Score: 0.9630
Training time:             0.18957710 seconds
Validation prediction:     0.00033980 seconds
Test prediction:           0.00029560 seconds
Samples per second (val): 82401 samples/sec
Samples per second (test): 94723 samples/sec
Samples per second (total): 88133 samples/sec
```





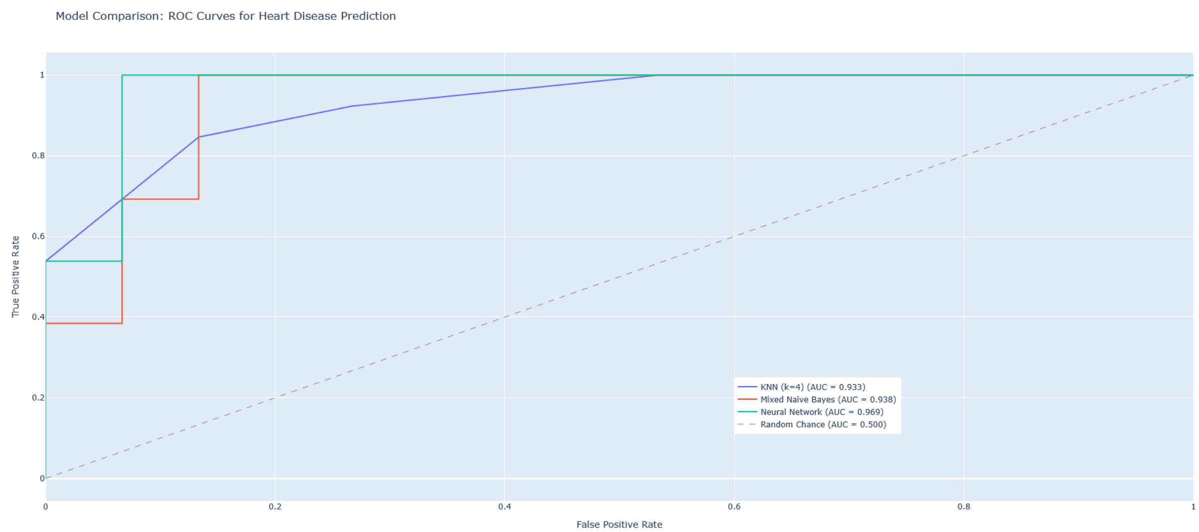
The Neural Network demonstrated exceptional performance across every metric, making it the most reliable and accurate model for this task.

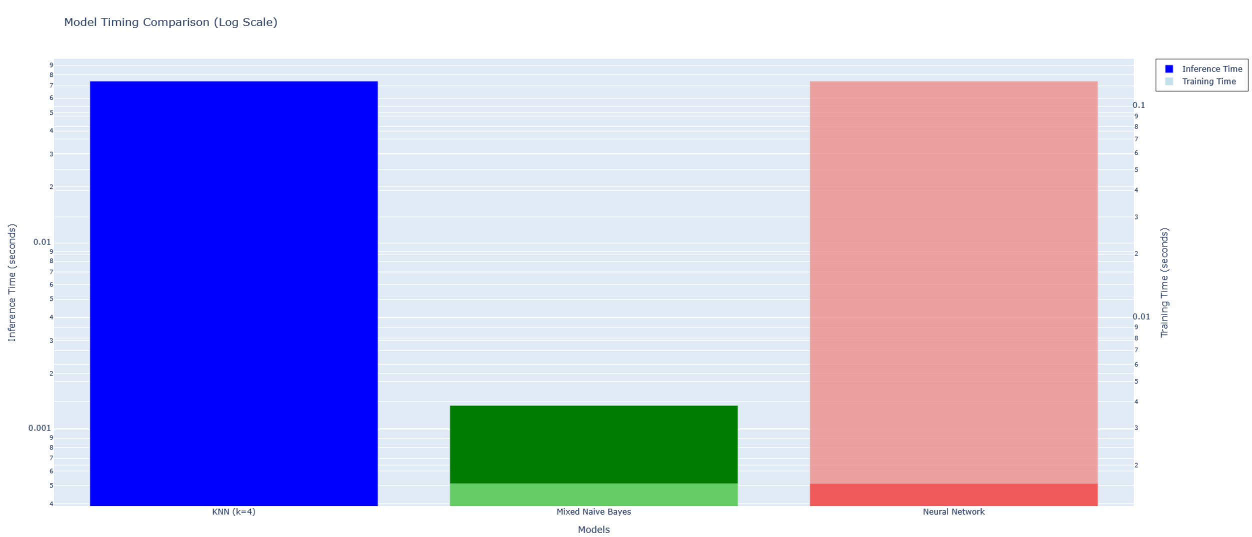
It achieved the best F1-Score (0.9630) and AUC (0.97), indicating superior overall performance and class discrimination. With a recall of 1.0, it successfully identified every single patient at risk of a heart attack in the test set (zero false negatives), which is crucial in the context of a medical diagnostic tool. Its high precision of 0.929 means it generated very few false alarms (only one false positive), ensuring that resources are not wasted and patients are not unnecessarily alarmed. It performed even better on the test set than on the validation set, proving it is extremely robust and generalisable, and not overfitted. It is the fastest model for making predictions, with the lowest inference time.

However, it has the highest training time of the three models, so it may require more time upfront before it can start making inferences. Second, neural networks, particularly deep ones, are black boxes; It can be difficult to understand why the model made a particular prediction, so clinicians may not be able to understand the reasoning behind a prediction to verify it and explain it to patients. Another challenge is ensuring that the model does not overfit. Neural networks, especially more complex ones, are prone to overfitting and memorising training data rather than learning patterns. Techniques such as regularisation and early stopping may be required to maintain generalisability.

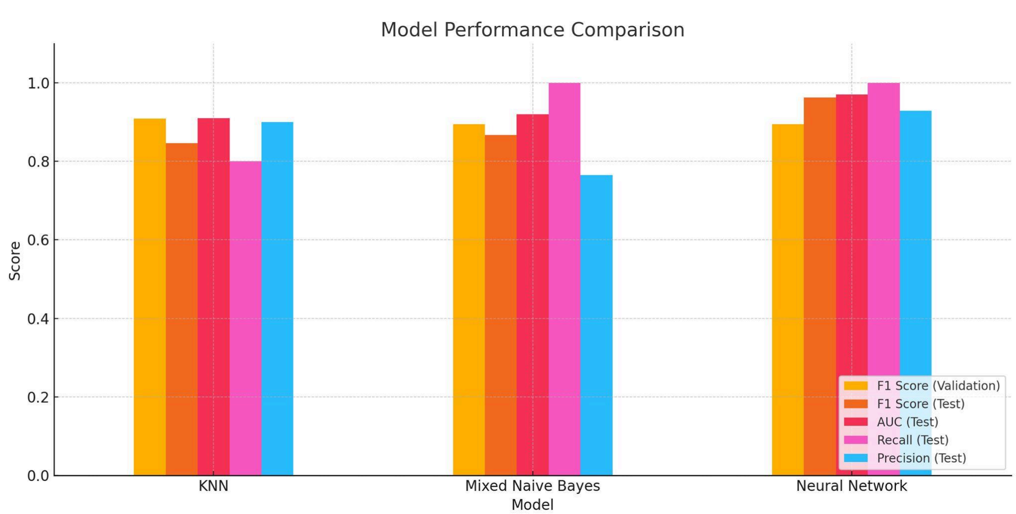
Comparison

Performance Evaluation





(Darker colour indicates inference time, lighter colour indicates training time)



Metric	**K-Nearest Neighbours (KNN)**	**Mixed Naive Bayes (MNB)**	**Neural Network (NN)**
F1 Score (Validation)	0.9091	0.8947	0.8947
F1 Score (Test)	0.8462	0.8667	**0.9630**
AUC (Test)	0.91	0.92	**0.97**
Recall (Test)	0.80	**1.00**	**1.00**
Precision (Test)	0.90	0.765	**0.929**
Inference Time	Slowest	Fast	**Fastest**
Training Time	Instantaneous	Very Fast	**Highest**
Generalisation	Moderate	Excellent	**Excellent**
Interpretability	Transparent	Good	Black-Box (none)

Neural Network

- *Highest F1-Score* (0.9630): Demonstrates the best balance between precision and recall
- *Highest AUC* (0.97): Excellent class discrimination capability
- *Perfect Recall* (1.0): Successfully identified all at-risk patients (zero false negatives)
- *Highest Precision* (0.929): Minimal false alarms with only 1 false positive
- *Exceptional Generalisation:* Performance improved from validation to test set, indicating strong robustness.

Mixed Naive Bayes

- *Perfect Safety Profile*: Recall of 1.0 ensures no at-risk patients are missed
- *Stable Performance*: Negligible drop between validation and test sets
- *Lower Precision* (0.765): Higher false positive rate (4 patients) compared to Neural Network
- *Good Generalisation*: Strong performance consistency across datasets

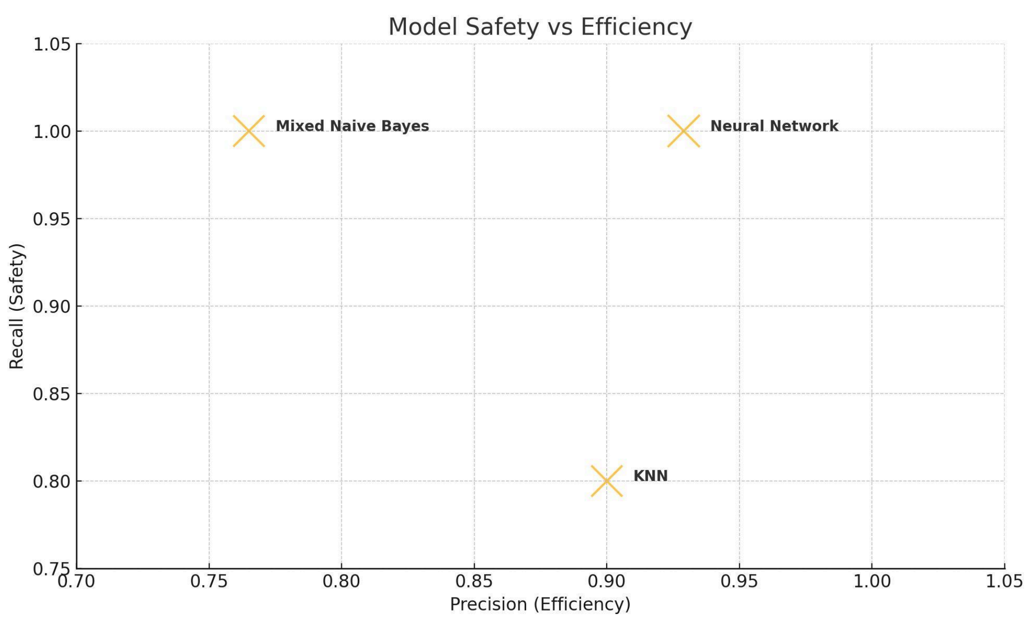
KNN with Gower Distance

- *Moderate Performance*: F1-score of 0.8462 and AUC of 0.91 show decent predictive power
- *Critical Safety Gap*: 2 false negatives represent unacceptable risk in medical context
- *Generalisation Issues*: Significant performance drop from validation to test set
- *Balanced Precision-Recall*: Balanced precision & recall

Timing

1. *Mixed Naive Bayes*: Fastest overall workflow with quick training and inference
2. *Neural Network*: Higher upfront training cost but fastest inference for production use
3. *KNN*: No training required but inference is orders of magnitude slower than other models

Evaluation in Medical Context



1. *Neural Network*: Ideal safety profile with zero false negatives and minimal false positives
2. *Mixed Naive Bayes*: Safe with zero false negatives but higher false positive rate

3. *KNN*: Unacceptable for medical use due to missed at-risk patients (false negatives)

Strengths & Weaknesses

Neural Network

Strengths: Exceptional overall performance with perfect recall and high precision, fastest inference time, and strong generalisation. Ideal balance of safety and efficiency for medical applications.

Weaknesses: Longest training time, "black box" interpretability issues that may concern clinicians, and potential susceptibility to overfitting without proper regularisation.

Mixed Naive Bayes

Strengths: Perfect safety profile with zero false negatives, fastest overall workflow, excellent stability across datasets, and highly interpretable results that clinicians can easily understand and trust.

Weaknesses: Higher false positive rate leading to unnecessary patient anxiety and resource allocation, and strong distributional assumptions that may not hold for all datasets.

KNN with Gower Distance

Strengths: No training phase required, handles mixed data types effectively through Gower distance, and provides intuitive classification logic based on similarity to neighboring cases.

Weaknesses: Unacceptable safety risk with false negatives, extremely slow inference time unsuitable for clinical use, poor generalisation, and high sensitivity to outliers and irrelevant features.

Potential Improvements

Neural Network

- *Interpretability*: Implement SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to provide feature importance explanations for clinical decision support
- *Regularisation*: Add dropout layers, L1/L2 regularisation, or batch normalisation to prevent overfitting and improve generalisation
- *Ensemble Methods*: Combine multiple neural networks with different architectures or training procedures to reduce prediction variance

Mixed Naive Bayes

- *Threshold optimisation*: Fine-tune classification thresholds to further reduce false positives while maintaining perfect recall.
- *Feature Engineering*: Create interaction terms or polynomial features to capture non-linear relationships while maintaining interpretability
- *Advanced Smoothing*: Implement adaptive smoothing techniques that vary based on feature characteristics rather than using uniform smoothing.

KNN

- *Advanced Distance Metrics*: Experiment with Mahalanobis distance or learned distance metrics to better capture feature relationships

- *Approximate Nearest Neighbors*: Implement algorithms like LSH (Locality-Sensitive Hashing) or Annoy to dramatically reduce inference time
- *Feature Selection*: Use techniques like Recursive Feature Elimination to identify the most relevant features for distance calculations
- *Weighted Voting*: Implement distance-weighted voting where closer neighbors have more influence on classification decisions

Conclusion

This comprehensive evaluation of three machine learning approaches for heart attack prediction demonstrates the critical importance of model selection in medical applications. While all three models showed competent performance, their suitability for clinical deployment varies significantly based on safety, accuracy, and practical considerations.

The Neural Network emerged as the superior solution, achieving exceptional performance with an F1-score of 0.9630, AUC of 0.97, and most importantly, perfect recall ensuring no at-risk patients were missed. Its combination of high precision (0.929) and fast inference time makes it ideal for real-world clinical deployment where both accuracy and speed are essential.

Mixed Naive Bayes proved to be a reliable safety-focused alternative, maintaining perfect recall while offering the advantage of interpretability that clinicians value. Although it generated more false positives, its transparency and consistent performance across validation and test sets make it a viable option for resource-constrained environments or applications where model explainability is paramount.

The KNN classifier, despite its intuitive approach and ability to handle mixed data types, demonstrated critical limitations for medical use. The presence of false negatives and extremely slow inference time render it unsuitable for clinical applications where patient safety and timely decision-making are non-negotiable.