Machine Learning Applications for Health
Predicting Heart Failure in Pneumonia Patients

Group 33

Abhishek Tummalapalli 1066956
Yusuf Berdan Guzel 1051639
Sayan Bachhar 1413063
Fangzhou Wang 1339824

November 2024

Contents

Abstract

Objective

This study aimed to use machine learning models to predict the likelihood of adult ICU patients diagnosed with pneumonia developing heart failure, within six months of their diagnosis.

Materials and Methods

Data was obtained from the MIMIC-IV database[1,2], starting with an initial sample of 21,846 patients with pneumonia. After data cleaning, the final cohort consisted of 8,677 pneumonia patients out of which 3,097 patients had heart failure. The dataset included patient information such as demographics, comorbidities, vital signs, medications, and laboratory results. Data preprocessing steps included processing missing values, normalising numerical features, and performing feature engineering. Using this, six models - Logistic Regression, Decision Tree, Random Forest, AdaBoost, XGBoost and Support Vector Classifier - were trained, optimised and compared across different performance metrics.

Results

The Random Forest model demonstrated the highest performance among all the six models. It initially attained 78.78% accuracy and an AUC of 0.8517, surpassing other models. Performing hyperparameter tuning improved it even further. It achieved an accuracy of 79.15% and an AUC of 0.8587 on the held-out test set, reinforcing its predictive power.

Discussion

The optimised Random Forest model demonstrated strong discriminatory ability in predicting heart failure in ICU pneumonia patients. SHAP value analysis revealed that key factors - such as creatinine, age, chronic kidney disease, diuretics and hyperlipidemia - are strong indicators to detect heart failure risk. Further analysis was performed across different demographics like gender and age, and the differences was discussed.

Conclusion

This study shows that machine learning models, particularly Random Forest, can accurately predict heart failure in ICU patients with pneumonia. This predictive approach could enable timely interventions, potentially improving outcomes in ICU care.

Introduction

Heart failure is a serious complication that poses significant risks to patients, especially in the intensive care unit (ICU) setting. This condition arises when the heart cannot pump blood efficiently, leading to decreased oxygen supply to vital organs, further worsening the patient's health and increasing the likelihood of hospital readmission[3]. The risk of heart failure is particularly elevated after pneumonia; studies have shown that inflammation and infection caused by pneumonia can exacerbate cardiovascular disease, potentially triggering or worsening heart failure[4]. Early identification of patients at high risk of heart failure following a pneumonia diagnosis can provide clinicians with valuable information to intervene promptly and improve patient prognosis.

In this study, we aim to predict the likelihood of heart failure development within six months of a pneumonia diagnosis in adult ICU patients. Recognising that both the direct effects of pneumonia and patients' pre-existing conditions significantly contribute to the risk of heart failure[5], we explore how factors like age, gender, vital signs, and comorbidities interact to influence this risk. Previous research has highlighted that advanced age, hypertension, diabetes, and chronic kidney disease are key predictors of cardiovascular complications, making these critical factors for analysis[6]. By leveraging large-scale clinical data, we seek to understand the complex relationships between these variables and heart failure outcomes to generate a predictive model.

To effectively analyse these multifaceted interactions, we employ machine learning techniques, which have emerged as powerful tools for clinical prediction. These methods allow for the examination of complex patterns that may not be immediately evident through traditional statistical approaches[7]. In the high-stakes environment of the ICU, improving the accuracy and timeliness of predictions is particularly valuable[8]. Therefore, we apply a range of machine learning models—from simple, interpretable models like Logistic Regression and Decision Trees to more sophisticated models such as Random Forests and XGBoost. This array of models offers a balance between interpretability and predictive power, essential for making informed medical decisions. By integrating machine learning into our analysis, this study not only aims to identify the best predictive approach but also to highlight the most important risk factors contributing to heart failure in pneumonia patients.

Methods

Data Collection

This study utilised data from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database to assess the likelihood of adult intensive care unit (ICU) patients with pneumonia developing heart failure within six months of their diagnosis. The MIMIC-IV database was selected due to its comprehensive clinical data and acceptance in medical research. An initial cohort of 21,846 patients was identified. To focus on individuals with potential follow-up data, the cohort was restricted to patients who had at least one more admission within the subsequent six months. After data cleaning to remove incomplete or inconsistent records, the final cohort consisted of 8,677 patients. Among these, 3,097 were diagnosed with heart failure, while 5,580 were not.

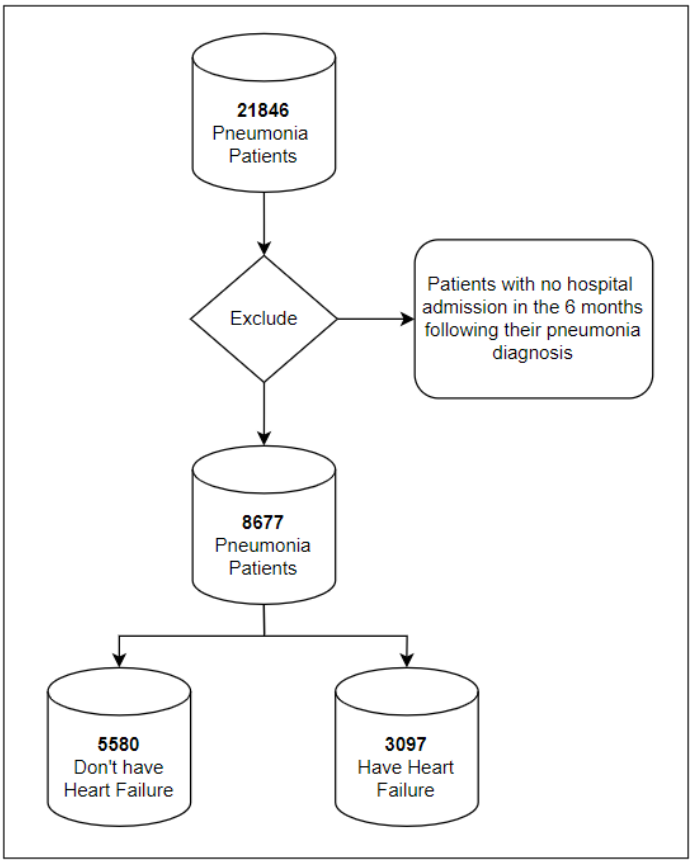

Figure 1. Pneumonia Patients Cohort Selection

Pneumonia diagnoses were identified using International Classification of Diseases, Ninth Revision (ICD-9) codes 480 to 486 and Tenth Revision (ICD-10) codes J12 to J18. Heart failure outcomes were tracked using ICD-9 code 428 and ICD-10 code I50. These specific ICD codes were selected as they correspond to general pneumonia and non heart transplant induced heart failure diagnoses, allowing accurate identification of relevant cases.

Demographic data collected included age and gender. Age is a known predictor for both pneumonia severity and heart failure complications[9], and gender differences have been observed in heart failure incidence following pneumonia, with males exhibiting higher risk[10]. Including these variables allowed for adjustment of potential confounding factors related to demographics.

Comorbidities such as hypertension[11], chronic kidney disease (CKD)[12], diabetes[13], and obesity[14] were included due to their significant association with increased risk of heart failure. These conditions can increase cardiac workload or impair cardiac function, contributing to the development of heart failure. Vital signs—including heart rate[14], blood pressure[11], and oxygen saturation (SpO_2)[15]—were collected as they are indicators of cardiovascular and respiratory status. Medications such as angiotensin-converting enzyme (ACE) inhibitors[16], beta-blockers[16], and diuretics were tracked due to their relevance in managing cardiovascular conditions and their potential influence on heart failure outcomes. Laboratory results, including troponin, creatinine, and bilirubin levels, were included as they provide insights into cardiac and renal function, both related to heart failure risk.

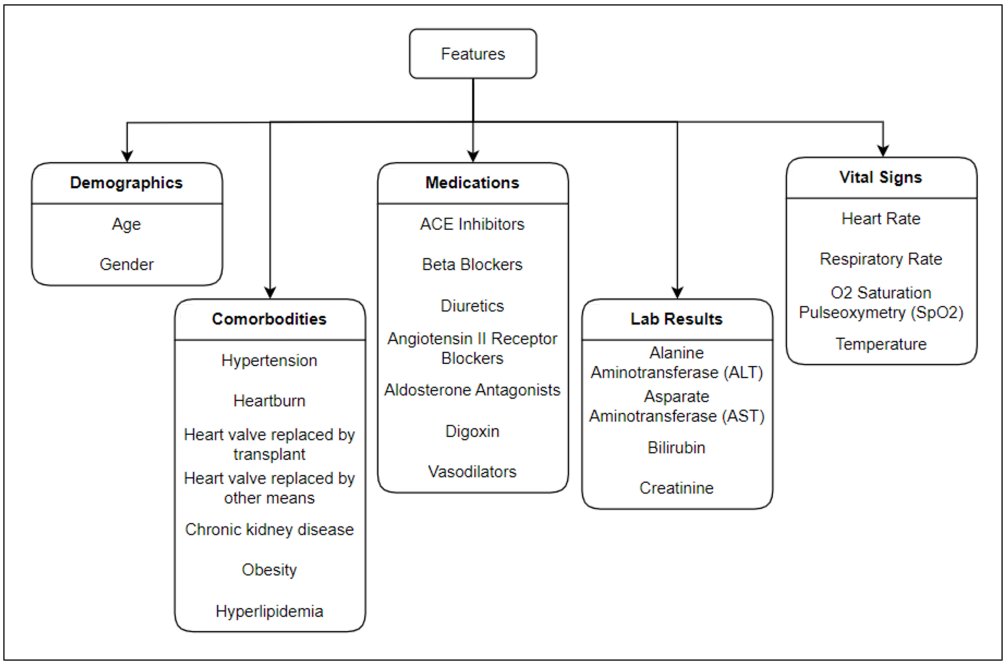

Figure 2. Features Related to Patient's Digital Phenotype

Data Preprocessing

Data preprocessing involved several steps to prepare the dataset for machine learning analysis. Columns with more than 50% missing values were excluded to increase data quality and reduce noise. Retaining columns with excessive missing data could introduce bias or unreliable patterns into the models. For numerical variables, missing values were imputed using the median to reduce the impact of outliers. For binary variables such as comorbidities and medications, missing values were treated as absent, under the assumption that non-documented conditions or treatments were not present.

Scaling of numerical features was performed to standardise the range of independent variables, which is essential for algorithms sensitive to feature scaling, such as Support Vector Classifier (SVC). This step prevents variables with larger scales from dominating the learning process.

Feature engineering was conducted to structure the data. Binary variables were created to represent the presence or absence of comorbidities and medications, to simplify categorical data into a suitable format. For vital signs and laboratory results, the mean, maximum, and minimum values over the patient's stay were calculated. These statistical measures capture the central tendency and variability of physiological parameters. This provided insights into patient health trends over time. Combining these temporal dynamics allows the model to assess fluctuations that could be indicative of deteriorating or improving conditions related to heart failure risk.

Model Selection

A comprehensive evaluation of six machine learning models was conducted to identify the most effective predictive model. Logistic regression was chosen as a baseline due to its simplicity and interpretability, providing a point of reference for more complex models. Decision Trees were included as a white-box model offering clarity and transparency in decision-making processes, which is valuable for understanding feature importance and model reasoning.

To increase predictive accuracy, several black-box models were employed, including Random Forest, AdaBoost, XGBoost, and Support Vector Classifier. These models are known for their ability to capture complex nonlinear relationships within data. The combination of both interpretable and sophisticated models allowed for a balance between transparency and performance.

To assess the models against simple heuristics, two baseline classifiers were included: a Random classifier that assigns labels with a 50-50 probability, and a Maximum classifier that always predicts the majority class. Including these baselines provided context for the performance of the more advanced models, benchmarking any improvements in predictive accuracy were meaningful.

Model Training and Optimisation

The dataset was split into training and testing sets using an 80/20 ratio. The training set was utilised for cross-validation to identify optimal hyperparameters and to evaluate each model's performance during development. GridSearchCV was employed for hyperparameter tuning, allowing systematic exploration of parameter combinations to improve model performance.

For the Random Forest model, hyperparameters such as the number of trees, maximum depth, and minimum samples per split were optimised. These parameters influence the complexity and generalisation ability of the model. For the Support Vector Classifier, the kernel type and regularisation parameter were tuned to adjust the model's flexibility and avoid overfitting.

Feature selection was conducted using Recursive Feature Elimination (RFE). RFE iteratively removes less important features based on model coefficients, allowing the identification of the most predictive variables. Subsets of 5, 10, 15, and 20 features were assessed to determine the optimal number of features that balance model complexity and predictive power.

Ethics Statement

The data for this project were sourced from the publicly available MIMIC-IV database, a widely used resource in medical research. Access to the data required completion of a certified training course, ensuring that users are equipped to handle the data responsibly. The dataset is de-identified in compliance with HIPAA standards, with robust privacy protections managed by the data custodians. These custodians implement strict protocols, inherently safeguarding sensitive information and providing a secure foundation for research. Given these comprehensive protections, no additional sensitivity measures were required for this project.

There are also no plans to redistribute any data generated during the analysis, as all work followed MIMIC-IV's guidelines for access and use. This approach not only ensures ethical handling of sensitive medical data but also aligns with the responsible use principles set forth by MIMIC-IV, maintaining data integrity and privacy throughout the research process.

Results

Model Performance

The performance of the initial set of machine learning models was evaluated using key metrics: Accuracy, Area Under the Receiver Operating Characteristic Curve (AUC), Precision, Recall, and F1-Score. Table 1 summarises the performance metrics for each model.

Table 1. Initial Performance Metrics of Machine Learning Models

Model	Accuracy	AUC	Precision	Recall	F1-Score
Random Classifier	0.5000	0.5000	0.3549	0.5000	0.4151
Maximum Classifier	0.6431	0.5000	-	0.0000	0.0000
Logistic Regression	0.7436	0.8020	0.6701	0.5543	0.6067
Decision Tree	0.7098	0.6791	0.5941	0.5902	0.5921
Random Forest	0.7878	0.8517	0.7379	0.6286	0.6789
AdaBoost	0.7518	0.8107	0.6717	0.5955	0.6313
XGBoost	0.7696	0.8337	0.6986	0.6233	0.6588
Support Vector Classifier (SVC)	0.7431	0.8002	0.6670	0.5595	0.6086

Among the evaluated models, the Random Forest algorithm exhibited the highest Accuracy (78.78%) and AUC (0.8517), indicating strong overall performance. Logistic Regression, AdaBoost, XGBoost and Support Vector Classifier also demonstrated favourable results, with AUC values exceeding 0.80.

Feature Selection and Hyperparameter Tuning

Feature selection was conducted using Recursive Feature Elimination (RFE) to identify the most predictive variables. The analysis indicated that retaining a larger number of features substantially improved model performance. Consequently, the decision was made to utilise all available features to maximise predictive accuracy.

Subsequent hyperparameter tuning was performed for all models. Table 2 summarises the tuned performance metrics for each model.

Table 2. Performance Metrics of Machine Learning Models after Hyperparameter Tuning

Model	Accuracy	AUC	Precision	Recall	F1-Score
Logistic Regression	0.7437	0.8019	0.6704	0.5543	0.6069
Decision Tree	0.7423	0.7994	0.6568	0.5818	0.6170
Random Forest	0.7908	0.8541	0.7392	0.6395	0.6857
AdaBoost	0.7571	0.8171	0.6844	0.5927	0.6352
XGBoost	0.7813	0.8453	0.7195	0.6346	0.6744
Support Vector Classifier (SVC)	0.7620	0.8077	0.7074	0.5680	0.6301

This process resulted in minor enhancements in performance metrics; however, the relative ranking of the models remained consistent. Random Forest maintained its position as the top-performing model across all evaluation metrics.

Optimised Model Performance on Test Dataset

The Random Forest model underwent further optimisation through hyperparameter tuning using GridSearchCV. The optimal hyperparameters identified were as follows:

Table 3. Optimised Random Forest Model Parameters

Bootstrap	Class Weight	Criterion	Max Depth	Max Features	Number of Estimators
False	Balanced	Entropy	100	`sqrt`	1000

The optimised Random Forest model was then evaluated on the 20% held-out test dataset. The performance metrics are presented in Table 4.

Table 4. Performance Metrics of the Optimised Random Forest Model on Test Dataset

Model	Accuracy	AUC	Precision	Recall	F1-Score
Random Forest (Optimised)	0.7915	0.8587	0.7471	0.6290	0.6830

Pareto Optimality Comparison

A Pareto optimality analysis was conducted to identify models that achieve the most favourable trade-offs across multiple performance metrics, including Accuracy, AUC, Precision, Recall, and F1-Score. Based on the evaluation metrics presented in Tables 1, 2 and 4, the optimised Random Forest model emerged as the sole Pareto optimal model. This model outperforms all other models across every metric, indicating that no other model provides a superior or equivalent performance in one metric without compromising in another. Consequently, the optimised Random Forest model offers the most balanced and effective performance for predicting heart failure in ICU patients with pneumonia, making it the optimal choice among the evaluated models.

Discussion

The optimised Random Forest model achieved an accuracy of 79.15% on the test dataset, accompanied by AUC score of 0.8587. These metrics indicate a strong discriminative ability of the model in differentiating between ICU pneumonia patients who will develop heart failure within six months and those who will not. The high AUC value, in particular, demonstrates the model's effectiveness in balancing sensitivity and specificity. These metrics are crucial in clinical settings where false positives and negatives can have significant implications for patient care.

Despite its performance, the Random Forest model has reduced interpretability compared to simpler models like Logistic Regression. Logistic Regression provides transparent coefficients that delineate the relationship between each predictor and the outcome. Interpretability makes it easier for clinicians to understand the underlying logic. In contrast, Random Forest operates as an ensemble of decision trees, making its decision-making process less transparent. This trade-off between performance and interpretability is significant in healthcare applications, since understanding the rationale behind predictions is important for clinical decision-making and gaining practitioner trust. Therefore, while Random Forest offers superior predictive capabilities, its complexity may pose challenges for clinical adoption unless supplemented with interpretability.

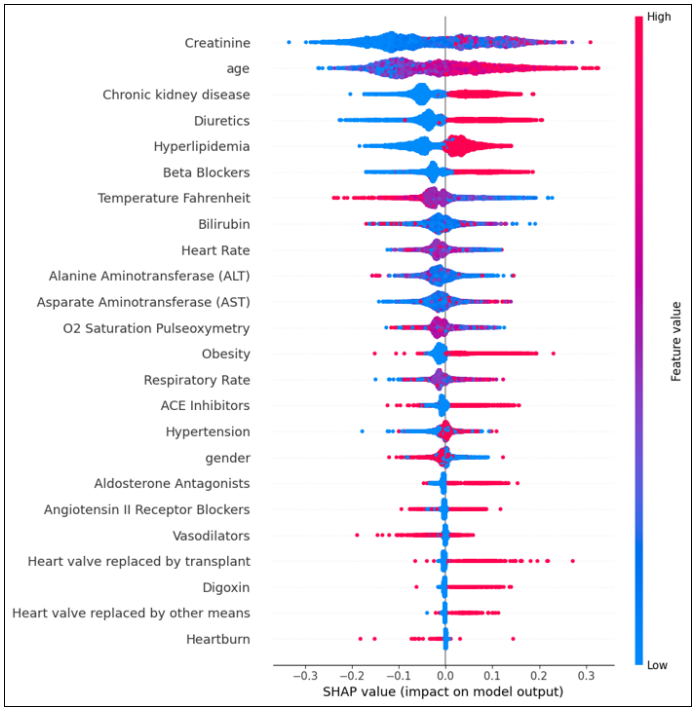

Figure 3. Feature Impact on Model Output

Interpretation[17]*: Positive SHAP value signifies increase in prediction of heart failure. Red color means higher feature values (1 for binary, big value for numeric), opposite for blue. A section of curve being thicker means more presence of the same feature value in the overall data. Features are ordered based on highest to lowest impact.*

Analysis of the plot in Fig. 3 revealed the key drivers of the model's predictions. The top five features identified were: Creatinine, Age, Chronic Kidney Disease, Diuretics and Hyperlipidemia. Creatinine and Chronic Kidney Disease highlighted the impact of renal function on cardiovascular health, while Age correlates with increased risk due to reduced physiological resilience. Diuretics, a treatment for fluid management, indicated the influence of medications on patient outcomes. Hyperlipidemia, tied to elevated blood lipids, reflected the model's focus on metabolic risk.

Expanding to the top ten features introduced Beta Blockers, Temperature, Bilirubin, Heart Rate, and ALT. These included vital signs and medications, signaling the model's ability to integrate diverse clinical factors. The presence of laboratory results, comorbidities, vital signs, and medications suggested that comprehensive clinical data can provide valuable insights for identifying high-risk individuals and guiding early interventions.

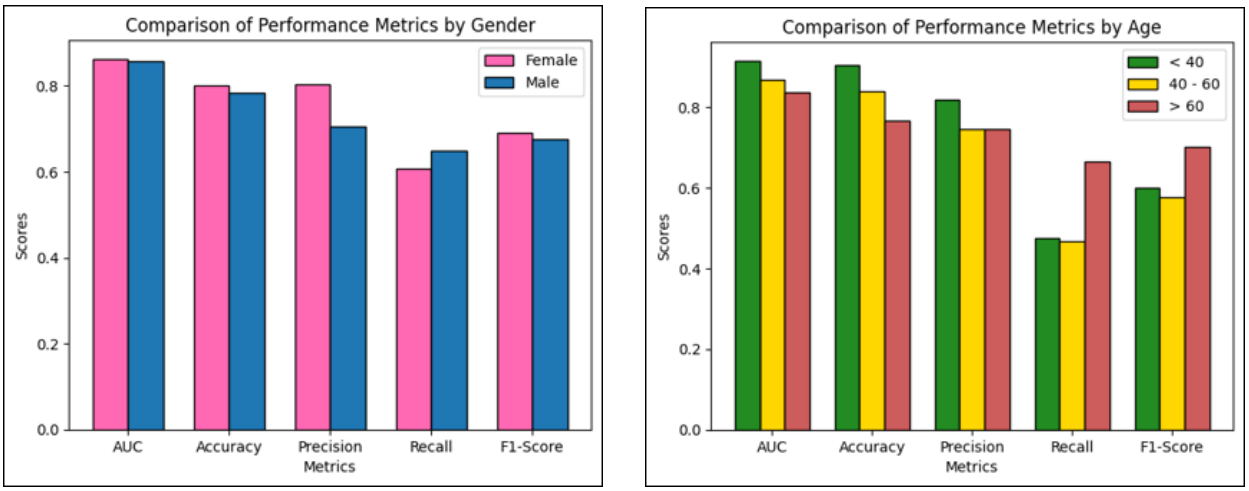

Figure 4. Comparison of Performance Metrics across Different Demographics

The model's performance exhibited variability across different demographic subgroups, particularly concerning gender. It demonstrated higher overall accuracy and marginally better precision in predictions for female patients. However, it also missed a significant number of positive cases among females, indicating a potential limitation in sensitivity for this group. Conversely, for male patients, the model showed improved sensitivity, effectively identifying a greater proportion of true positive cases, albeit with slightly lower overall accuracy compared to females. This gender-based disparity may stem from underlying biological differences in how heart failure manifests and progresses between males and females, as well as variations in comorbidity profiles and treatment responses. Such differences highlight the necessity for models to account for gender-specific factors to enhance predictive reliability and equity in clinical settings.

Age stratification further showcased the model's performance dynamics. The highest accuracy was observed in the youngest age group, suggesting that the model effectively identifies heart failure risk in younger patients. However, it struggled to detect a considerable number of positive cases within this cohort, indicating lower sensitivity. In contrast, the model's ability to identify true positive cases improved with increasing age, particularly in the oldest age brackets. This trend may reflect the greater prevalence and complexity of comorbid conditions in older populations, providing more distinctive patterns for the model to recognise. Nonetheless, the slight decrease in overall accuracy and reliability in the oldest age groups suggested that while the model was adept at recognising high-risk individuals, there is room for improvement in consistently identifying true positives across all age ranges. This could be addressed by incorporating age-specific variables or adjusting model parameters to enhance performance in younger populations.

The predominance of diverse clinical factors among the top predictive features shows the importance of a holistic approach to monitoring in the early detection and management of heart failure. These features include patient demographics, laboratory biomarkers, comorbidities, vital signs and medications, each reflecting different aspects of patient health. By capturing not only underlying physiological states but also the effects of treatments and broader health conditions, the model provides a comprehensive view of the factors influencing heart failure risk. Integrating these diverse elements into predictive models enhances their ability to detect subtle changes that may precede clinical symptoms. This makes the model a valuable tool for clinicians, enabling proactive management and timely interventions to reduce the likelihood of adverse outcomes.

Conclusion

This study successfully developed and optimised a Random Forest model to predict the onset of heart failure in adult ICU patients with pneumonia within six months of admission, achieving an accuracy of 79.15% and AUC score of 0.8587. These results demonstrate the model's strong predictive capability, particularly highlighting the importance of different clinical factors in forecasting heart failure risk. However, the study is subject to several limitations. The reliance on the MIMIC-IV database may constrain the generalisability of the findings to other clinical settings with different patient populations and data characteristics. Additionally, the inherent complexity of the Random Forest model reduces its interpretability, posing challenges for clinical adoption where transparency is crucial for decision-making.

Future work should focus on enhancing model interpretability through the integration of explainable AI techniques, which can provide clinicians with clearer insights into the model's decision-making process. Furthermore, expanding the dataset to include diverse populations and incorporating additional clinical variables could improve the model's robustness and applicability across various healthcare environments. While the optimised Random Forest model presents a promising tool for early identification of heart failure risk in ICU pneumonia patients, addressing its limitations is essential to maximise its clinical utility and ensure equitable healthcare delivery.

References

1. Alistair E. W. Johnson, Tom J. Pollard, Lucas A. Celi, Leo Anthony Celi, and Roger G. Mark. MIMIC-IV (version 2.2). https://doi.org/10.13026/6mm1-ek67, 2023.

2. Alistair E. W. Johnson, Leo Bulgarelli, Lu Shen, and et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023.

3. Omar Dar and Martin R. Cowie. Acute heart failure in the intensive care unit: Epidemiology. *Critical Care Medicine*, 36(1):S3, 2008.

4. Vicente F. Corrales-Medina, Katherine N. Alvarez, Lisa A. Weissfeld, Derek C. Angus, Julio A. Chirinos, Charles C.H. Chang, Anne B. Newman, Laura Loehr, Aaron R. Folsom, Mitchell S.V. Elkind, and Matthew F. Lyles. Association between hospitalization for pneumonia and subsequent risk of cardiovascular disease. *JAMA*, 305(1):63–69, 2011.

5. Lu Shen, Pardeep S Jhund, Inder S Anand, Akshay S Bhatt, Akshay S Desai, Aldo P Maggioni, Fábio A Martinez, Marc A Pfeffer, Amir R Rizkala, Jean L Rouleau, Karl Swedberg, Muthiah Vaduganathan, Orly Vardeny, Dirk J van Veldhuisen, Faiez Zannad, Michael R Zile, Milton Packer, Scott D Solomon, and John JV McMurray. Incidence and outcomes of pneumonia in patients with heart failure. *Journal of the American College of Cardiology*, 77(16):1961–1973, 2021.

6. Connie W. Tsao, Aaron W. Aday, Zaid I. Almarzooq, Alvaro Alonso, Angela Z. Beaton, Marcio S. Bittencourt, Amelia K. Boehme, Alfred E. Buxton, April P. Carson, and Yvonne Commodore-Mensah. Heart disease and stroke statistics—2021 update: A report from the american heart association. *Circulation*, 143(8):e254–e743, 2021.

7. Clare R Olsen, Robert J Mentz, Kevin J Anstrom, David Page, and Parag A Patel. Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure. *American Heart Journal*, 229:1–17, 2020.

8. Benjamin Shickel, Patrick J. Tighe, Azra Bihorac, and Pegah Rashidi. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604, 2017.

9. American Heart Association. Heart failure: Understanding the condition. https://www.heart.org/en/health-topics/heart-failure, 2021.

10. C. W. Tsao, A. W. Aday, Z. I. Almarzooq, A. Alonso, A. Z. Beaton, M. S. Bittencourt, A. K. Boehme, A. E. Buxton, A. P. Carson, and Y. Commodore-Mensah. Heart disease and stroke statistics—2021 update: A report from the american heart association. *Circulation*, 143(8):e254–e743, 2021.

11. P. K. Whelton, R. M. Carey, W. S. Aronow, D. E. Casey, K. J. Collins, C. D. Himmelfarb, S. M. DePalma, and S. Gidding. 2017 acc/aha/aapa/abc/acpm/ags/apha/ash/aspc/nma/pcna guideline for the prevention, detection, evaluation, and management of high blood pressure in adults. *Hypertension*, 71(6):e13–e115, 2018.

12. Kidney Disease: Improving Global Outcomes (KDIGO). Kdigo 2020 clinical practice guideline for diabetes management in chronic kidney disease. *Kidney International Supplements*, 10(4):S1–S115, 2020.

13. American Diabetes Association. Standards of medical care in diabetes—2021. *Diabetes Care*, 44(Supplement 1):S1–S232, 2021.

14. G. N. Levine. Cholesterol lowering in the primary and secondary prevention of coronary heart disease. *The New England Journal of Medicine*, 336(4):262–273, 1997.

15. B. R. O'Driscoll, L. S. Howard, and J. Earis. Bts guideline for oxygen use in adults in healthcare and emergency settings. *Thorax*, 72(Supplement 1):i1–i90, 2017.

16. Michael H Strauss, Andrew S Hall, and Krzysztof Narkiewicz. The combination of beta-blockers and ace inhibitors across the spectrum of cardiovascular diseases. *Cardiovascular Drugs and Therapy*, 37(4):757–770, 2023.

17. Aidan Cooper. Explaining machine learning models: A non-technical guide to interpreting shap analyses. https://www.aidancooper.co.uk/a-non-technical-guide-to-interpreting-shap-analyses/, 2021.

Appendix

All the codes used to generate the results in this report are present in the following Github link: *COMP90089-Machine-Learning-Applications-for-Health.git*

It includes a README file explaining the content of the files as well as instructions on how to run them.

Table 5. Contributor Role Taxonomy (CRediT) Table

Role	**Contributors**
Conceptualisation	Yusuf Berdan Guzel
Data Curation	-
Formal Analysis	Fangzhou Wang, Sayan Bachhar, Yusuf Berdan Guzel
Funding Acquisition	-
Investigation	Fangzhou Wang, Sayan Bachhar, Yusuf Berdan Guzel
Methodology	Yusuf Berdan Guzel
Project Administration	Yusuf Berdan Guzel
Resources	-
Software	-
Supervision	-
Validation	Sayan Bachhar, Yusuf Berdan Guzel
Visualisation	Sayan Bachhar
Writing – Original Draft	Abhishek Tummalapalli, Fangzhou Wang
Writing – Review & Editing	Sayan Bachhar, Yusuf Berdan Guzel