PREDICTING ICU MORTALITY FOR PATIENTS WITH MULTIPLE CHRONIC CONDITIONS

POLISLAVA KAMENOVA KALCHEVA

THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF SCIENCE IN COGNITIVE SCIENCE AND ARTIFICIAL INTELLIGENCE

AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES OF TILBURG UNIVERSITY

STUDENT NUMBER

2071490

COMMITTEE

dr. Phillip Brown

dr. Silvy Collin

LOCATION

Tilburg University

School of Humanities and Digital Sciences

Department of Cognitive Science & Artificial Intelligence

Tilburg, The Netherlands

DATE

May 16th, 2025

WORD COUNT

8797

PREDICTING ICU MORTALITY FOR PATIENTS WITH MULTIPLE CHRONIC CONDITIONS

Polislava Kamenova Kalcheva

Abstract

This thesis explores whether machine learning (ML) models predict intensive care unit (ICU) mortality more efficiently for admissions of patients with multiple chronic conditions (MCC=1) compared to those with only one chronic condition, along with possible non-chronic conditions (MCC=0). While previous research applied ML to the whole group of ICU patients and reported strong results (Nistal-Nuño, 2022; Gao et al., 2024), they have not assessed model behaviour independently for admissions in the MCC=1 group compared to those in the MCC=0 group.

Using the MIMIC-III Clinical Database, this thesis trained and assessed four ML models, Logistic Regression (LR), XGBoost, Random Forest (RF), and k-Nearest Neighbors (KNN), on the MCC groups. Precision, recall, and F1-score were on focus, as selected in other ML studies (Subudhi et al., 2021; Darabi et al., 2018), rather than accuracy and AUC.

After excluding the LR and KNN, which performed lower than XGBoost and RF, XGBoost and RF were compared. XGBoost outperformed RF in all metrics in MCC=1. XGBoost had a higher recall in MCC=0, which is most crucial in mortality prediction. Therefore, it was still preferred despite the lower precision and F1 than

RF.

DATA SOURCE, ETHICS, CODE AND TECHNOLOGY STATEMENT

Data Source

Who is the owner of the data? PhysioNet.org is the owner of the dataset. https://physionet.org/content/mimiciii/1.4/, corresponding author- Alistair Johnson, Massachusetts Institute of Technology alistairewj@gmail.com

Does the thesis project involve collecting data from human participants or animals? No

Does the owner of the data give consent? Yes.

How and through what channel is the data acquired? I had to apply through PhysioNet.org (Goldberger et al., 2000), and the requirement was to get certified for a specific course called 'Human Research - Data or Specimens Only Research' in the CITY Program website https://about.citiprogram.org/. Also, I had to give personal information and my reference information, who is the supervisor of my thesis. After that, I had to wait for them to contact my reference, to check my results from the training course, and then I got approval, and access to the dataset was opened for me. In the end, I was able to download the CSV files that I needed to write the code for my research.

Additional: The dataset includes different diseases. To extract only the chronic diseases, this website https://hcup-us.ahrq.gov/toolssoftware/chronic/chronic.jsp#files was used, providing a CCI2015.CSV file containing all general names and codes for each chronic disease known from general knowledge.

Figures

Did you create all the images and figures? Yes.

Code

Did you use parts of the code from another study/ someone else? No.

Did you list, including version number, all libraries and frameworks used? Yes.

Technology

Did you use any tools or services to paraphrase the given text? Thesaurus.com was used to include better English words in my report.

Did you use any tools or services to check spelling or grammar? Grammarly.com was used to correct any possible grammar mistakes in my report.

Did you use any tools or services to typeset the given text? No.

Did you use any reference management software other than the LaTeX template? Microsoft Word was used for writing this report.

Did you use any generative language models to help with your writing? No.

Source/Code/Ethics/Technology Statement

Data Source: The data used in this thesis is the MIMIC-III Clinical Database v1.4 https://physionet.org/content/mimiciii/1.4/, acquired via request from PhysioNet.org (Goldberger et al., 2000). The dataset is anonymized and publicly available. One of the requirements was completing the "Human Research - Data or Specimens Only Research" course from the CITI Program and receiving approval via PhysioNet.org (Goldberger et al., 2000). The dataset is available from PhysioNet, and the corresponding author is Alistair Johnson. No new data were collected from human participants or animals. To find out which are the chronic diseases among all different diseases in the MIMIC-III Clinical Database, I used the HCUP CCI tool (https://hcup-us.ahrq.gov/toolssoftware/chronic/), specifically the CCI2015.CSV file. All figures/tables in this thesis were created by the author. Code: All code was written by the author using Python. No external code was reused or adapted. A complete list of Python libraries is documented in section 3.7.2 of the thesis. Technology: The code for this thesis was created in Python 3.9.7, using Jupyter Notebook, Windows 11. No typesetting tools or generative AI models were used. Thesaurus.com and

Grammarly.com were used for word ideas and grammar checks, respectively. The thesis was written in Microsoft Word.

1. Introduction

1.1. Context and Goal

This research has a goal to explore whether machine learning (ML) models can accurately predict intensive care unit (ICU) mortality for admissions of patients with multiple chronic conditions (MCC) compared to general ICU admissions of patients who have one chronic condition, along with possible non-chronic conditions. Accurate mortality prediction in the ICU is significant for resource distribution and time-consuming interventions. Traditional scoring systems, such as SAPS II and SOFA, are frequently used in the ICU, but they have limitations in handling complex diseases and large-scale mixed data (Kong et al., 2020; Nistal-Nuño, 2022; Gao et al., 2024). Some studies have shown that ML algorithms can outperform traditional methods by taking examples from electronic health records (Subudhi et al., 2021; Darabi et al., 2018). Nonetheless, most previous research has not directly shown the differences in model performance between MCC admissions of patients and those with 1 chronic condition, along with possible non-chronic conditions.

1.2. Scientific Relevance

This research contributes to enhancing critical care by applying ML focusing on MCC admissions of patients, a belittled subgroup in many of the previous mortality prediction research studies. While ML models, such as XGBoost, Random Forest (RF), and Neural Networks (NN), have shown high efficiency in ICU mortality in patients in general (Iwase et al., 2022; Mamandipoor et al., 2021; Ye et al., 2020), none have shown performance in chronic condition ICU mortality. For example, there are AUC values of 0.91 using XGBoost in ICU patients in general, focusing on ML's predictive efficiency (Nistal-Nuño, 2022; Gao et al., 2024). This thesis focuses on evaluating how ML models' performance changes when focusing on MCC admissions of patients, who commonly show

more complex physiological profiles and a higher risk in making health worsening. Therefore, there is a gap in the performance analysis of ML models in ICU settings.

1.3. Societal Relevance

This research has clear societal value. MCC admissions of patients are more likely to suffer from injurious ICU outcomes and need more complex interventions (Alghatani et al., 2021; Mamandipoor et al., 2021). Healthcare systems face growing tension to provide customized, proper, low-cost, but effective care. So, predictive models that can indicate high-risk MCC admissions of patients may implement more focused clinical decision-making. For instance, a high recall metric in mortality prediction could indicate earlier interventions for patients likely to decline their health. Additionally, high precision metric helps in the prevention of overdoing intensive interventions. By enhancing mortality prediction for MCC admissions of patients, this research can support more unbiased and effective use of ICU resources.

1.4. Research Strategy and Approach

This study used the MIMIC-III database, an ICU dataset that is publicly available and used worldwide. It contains structured data from over 58,976 hospital admissions. Admissions of patients were separated into two groups: MCC=0, including admissions with one chronic condition and possibly non-chronic conditions, and MCC=1, including admissions with more than one chronic condition only. After preprocessing the raw data, four ML models were trained to predict binary mortality outcomes- Logistic Regression (LR), XGBoost, RF, and k-Nearest Neighbors (KNN). These models were included in the analysis, but the outputs of RF and XGBoost clearly outperformed those of LR and KNN. This was because of the ability of RF and XGBoost to work efficiently with feature selection and to capture complex relationships in the data. Evaluation metrics included accuracy, precision, recall, and F1-score, and AUC. In ICU mortality prediction, precision,

recall, and F1-score are significantly important. High recall identifies at-risk patients and reasonable precision avoids false alarms. Therefore, they can notably impact patient outcomes and ICU resource allocation. This thesis prioritizes recall, precision, and F1-score because they better show model performance in imbalanced ICU mortality prediction (Subudhi et al., 2021). Subudhi et al. (2021) demonstrated that AUC can cover low sensitivity, risking missed deaths. Moreover, Darabi et al. (2018) discovered that accuracy is untrustworthy in clinical imbalance and preferred recall and F1 instead. Analyses were made to compare model performance between the two patient groups (MCC=0 and MCC=1). Feature importance was applied using built-in functions provided by the models.

1.5. Research Question: Can machine learning accurately predict ICU mortality in patients with multiple chronic conditions compared to patients with one chronic condition, along with possible non-chronic conditions?

1.6. Findings

This study assessed the performance of four ML models, LR, XGBoost, RF, and KNN, for predicting ICU mortality among two patient groups, MCC=1 and MCC=0. The focus was placed on significant metrics for clinical prediction tasks such as precision, recall, and F1, as they are more suggestive and reliable of a model's ability to identify true mortality cases than accuracy or AUC alone.

LR and KNN performed overall lower than XGBoost and RF, so the evaluation continued with a focus on XGBoost and RF only.

Therefore, in the MCC=0 group, RF achieved slightly better overall balance with higher accuracy (0.95 vs. 0.92), precision (0.23 vs. 0.17), and F1 (0.32 vs. 0.27) compared to XGBoost. Nevertheless, XGBoost demonstrated higher recall (0.64 vs. 0.50) than RF, which is a crucial metric in mortality prediction, where finding all high-risk patients is crucial to prevent fatal outcomes. Although AUC was equal for both models (0.90) in the MCC=0

group, the higher recall of XGBoost makes it a stronger nominee for mortality prediction, where false negatives are particularly costly.

On the other hand, XGBoost outperformed Random Forest across all evaluation metrics for the MCC=1 group. It accomplished higher accuracy (0.73 vs. 0.70), precision (0.20 vs. 0.18), recall (0.68 vs. 0.67), F1-score (0.31 vs. 0.28), and AUC (0.79 vs. 0.76). This consecutive advantage accentuates XGBoost's greater ability to handle the raised complexity of patients with MCC, which is exceptionally valuable in the ICU, where delicate predictions are needed.

When comparing MCC=1 and MCC=0 groups using the XGBoost model, performance was better in the MCC=1 group for precision (0.20 vs. 0.17), recall (0.68 vs. 0.64), and F1-score (0.31 vs. 0.27), which are the most important clinical metrics in mortality prediction. Although the MCC=0 group had slightly higher accuracy and AUC, these are less impactful when dealing with imbalanced data and mortality risks.

Consequently, XGBoost is the most effective model overall, especially for the MCC=1 group. It outperforms RF, such as providing stronger predictive results than RF, which identifies high-risk admissions of patients more accurately. In the image below, both ML models are compared for each group.

Moreover, this thesis had the aim to evaluate whether the best-performing ML model, XGBoost, could outperform the traditional model, SAPS II score, considered one of the most accurate in ICU mortality prediction scores. SAPS II has been shown to predict ICU mortality with an AUC of 0.77, and it is suggested that ML models can marginally enhance on this benchmark (Kong et al., 2020). Therefore, an evaluation was performed by applying the XGBoost model to the whole group of admissions of patients from the MIMIC III clinical database, not limited to those with MCC=1 only. The XGBoost model achieved an AUC of 0.81, confirming the theory proposed by Kong et al. (2020) that ML models can

improve predictive performance beyond traditional tools.

Nonetheless, this evaluation was conducted on the general ICU population. So far, no prior studies have specifically compared traditional scoring systems and ML models within MCC=0 and MCC=1 subgroups. This thesis fills that gap by showing that for admissions of patients with MCC=1, ML, specifically XGBoost, outperforms RF and suggests better mortality prediction than traditional models.

2. Related Work

2.1. Area of Research, Research Issue, Issue Importance

This research is in the field of clinical predictive modelling, with a specifying on ICU mortality prediction using ML. The main challenge is to evaluate and enhance the capability of ML algorithms to predict the accuracy of mortality outcomes for admissions of patients in MCC=1 and MCC=0 groups, and to compare them to see whether the MCC=1 group needs prioritization in the ICU. While ICU mortality prediction has been researched a lot, the performance of ML models on admissions of patients in the MCC=1 group has not been evaluated in existing literature. This thesis reviews recent and relevant research work, prioritizing frequently used models, datasets, evaluation metrics, and how previous studies compare to and inform the current research.

2.2. Relevant Work in The Same Research Area, Relevant Theories, State of The Art

Multiple studies have compared ML algorithms with traditional ICU scoring systems such as SAPS II, SOFA, and APACHE II. Kong et al. (2020) showed that traditional models achieved balanced predictive performance. For example, SAPS II showed AUC = 0.77 in ICU mortality prediction. However, these models lacked adaptability in handling mixed patient populations. On the other hand, more recent research demonstrated that XGBoost had an AUC = 0.919, which is higher than SAPS II in ICU patient data (Nistal-Nuño, 2022). Moreover, static clinical scores in mortality prediction are consistently

outperformed by the RF model (Gao et al., 2024; Mamandipoor et al., 2021).

2.3. Used Models by Researchers, The State of The Art

Several studies used ML models that were applied in this thesis. For example, ICU mortality was evaluated by RF model and was found to be highly efficient in capturing complex feature interactions (Darabi et al., 2018). Moreover, a combination of RF and LR models was evaluated, confirming that tree-based methods are more suitable for ICU complex data including many features based on an evaluation of combination of RF and LR models (Ye et al., 2020). Therefore, LR is outperformed by non-linear models. Additionally, XGBoost's capability to model non-linearity and feature interactions makes it suitable for ICU outcome prediction (Iwase et al., 2022).

2.4. Research Gaps, Possible Alternative Approaches

No study focused on patients with multiple chronic conditions in the ICU. Most of the previous studies applied ML models to patients in the ICU in general. For example, the clinical declining risk of mortality was examined in general (Subudhi et al., 2021), but has not been evaluated in chronic condition subgroups. Moreover, AUC performance was evaluated, but the differences in outcomes in MCC subgroups were not examined (Gao et al., 2024). Additionally, there are factors influencing ICU mortality in COVID-19 patients, and complex conditions influence patients' health (Alghatani et al., 2021), but MCC as a predictive factor was not in focus. Therefore, the current thesis aims to fill this lack of analysis that leads to a crucial gap.

2.5. Limitations of Existing Models, Brought Insights by My Research, and Solving The Problems

For mortality prediction, traditional ICU scoring systems such as APACHE II, SAPS II, and SOFA have been used for a long time, but critical limitations are associated with them, specifically when applied to patients with MCC. These models rely on fixed scoring

rules based on defined thresholds and are unable to identify complex, nonlinear interactions between patient variables. For example, APACHE II has shown decreasing performance over time because of aging ICU populations and treatments (Iwase et al., 2022). SAPS II is based on older data and struggles with generalizability to modern ICU populations (Kong et al., 2020).

Moreover, for handling high-dimensional data or large electronic health record datasets efficiently, the traditional models were not designed to do so (Kong et al., 2020). SOFA is useful for organ dysfunction assessment and offers lower predictive accuracy than modern ML models. SOFA often needs more patient information to reach similar performance (Nistal-Nuño, 2022). Their static scores limit their adaptability to patient-specific risk profiles, especially in the presence of diseases like chronic conditions.

On the other hand, RF and XGBoost automatically capture non-linear relationships because these models can manage well with large and complex feature spaces and can be fine-tuned for subgrouping. For instance, ML models outperform traditional methods in predictive accuracy for general ICU populations (Subudhi et al., 2021; Gao et al., 2024). Nonetheless, previous studies have not focused on the subgroups of general ICU admissions of patients, MCC=0, and admissions of patients in the MCC=1 group. This forms a gap in whether ML models are equally effective, or whether prioritizing the higher-risk MCC=1 population is needed.

This research discusses that gap by comparing model performance across MCC=0 and MCC=1 subgroups. It finds the best-performing ML model for each subgroup and shows whether the chosen best-performing ML model improves the traditional scoring systems when comparing the whole group of admissions of patients, not only the MCC=0 or MCC=1.

2.6.Research Question and Goals of the Current Work of My Research

This research focuses on the question “Can machine learning accurately predict ICU mortality in patients with multiple chronic conditions compared to patients with one chronic condition, along with possible non-chronic conditions?”. This question shows a crucial gap in the current literature, where traditional scoring systems frequently cannot handle the complexity of patients with complex diseases. Previous ML studies have not focused on the performance of the chronic condition influence.

The main goal of this study is to assess and compare the performance of four ML models, LR, XGBoost, RF, and KNN, in predicting ICU mortality. The models are applied to two subgroups of admission of patients derived from the MIMIC-III dataset: MCC=0, including admissions of patients with exactly one chronic condition along with possible non-chronic conditions, and MCC=1, including admissions of patients with more than one chronic condition only. By evaluating the groups individually, the research has a goal to determine which and whether ML models are more efficient at finding mortality in the MCC=1 group, and whether ML models outperform traditional models such as SAPS II in mortality prediction.

Ideally, this research will find the most trustworthy ML model for ICU mortality prediction across both subgroups, with a focus on enhancing accuracy, recall, and F1. The findings will support the prioritization for admissions of patients in the MCC=1 group to prevent their mortality in the ICU.

2.7. Methodology of This Thesis

This research applies and compares four ML models, LR, XGBoost, RF, and KNN, to predict mortality in MCC=0 and MCC=1 groups. The models are using the precision, recall, F1-score, accuracy, and AUC metrics, but the focus is on precision, recall, and F1, which are more clinically meaningful for mortality prediction, as stated in Subudhi et al. (2021).

2.8. My Research Fills The Current Gaps, Dataset Description

This research fills a crucial gap in the previous studies on ICU mortality prediction. While many previous studies have strongly applied ML models to ICU data, they have treated ICU patients as a single general population. Until now, no studies have focused on patients in the ICU with chronic conditions. Especially, comparing those with MCC=0 to MCC=1. This lack of subgroup-specific evaluation leaves ambiguity in whether traditional models perform equally well for higher-risk, more complex patients as the patients with less complex health issues.

Therefore, this research introduces a new framework using the MIMIC III clinical database, containing detailed clinical information about 58,976 hospital admissions at the Beth Israel Deaconess Medical Center. Admissions of patients are separated into MCC=0 and MCC=1 groups. MCC=0 group includes admissions in ICU with one chronic condition along with possible non-chronic conditions, and MCC=1 group includes admissions in ICU with more than one chronic condition only.

By training and testing multiple ML models on these two groups individually and comparing their performance using the precision, recall, and F1 metrics, this thesis shows a new analysis of how ML models perform in different chronic condition subgroups. Therefore, this thesis identifies the XGBoost ML model as the most effective model for ICU mortality prediction and suggests new insights into how ML can better contribute to clinical decision-making in high-risk ICU subgroups.

2.9. Hypotheses

H_0 (Null Hypothesis): Machine learning models do not improve ICU mortality prediction accuracy for admissions of patients with multiple chronic conditions compared to admissions of patients with one chronic condition, along with possible non-chronic conditions.

H_1 (Alternative Hypothesis): Machine learning models improve ICU mortality prediction accuracy for admissions of patients with multiple chronic conditions compared to admissions of patients with one chronic condition, along with possible non-chronic conditions.

The null hypothesis (H_0) assumes that ML models do not enhance ICU mortality prediction for admissions of patients in the MCC=1 group compared to admissions of patients in the MCC=0 group. On the other hand, the alternative hypothesis (H_1) suggests that ML models do provide better predictive performance for the MCC=1 group compared to the MCC=0 group.

This assumption is supported by studies illustrating that ML models outperform traditional scoring systems, especially in high-risk cases. For example, Subudhi et al. (2021) and Mamandipoor et al. (2021) discovered that ML models performed well for patients with complex health conditions. Moreover, Gao et al. (2024) and Alghatani et al. (2021) showed that models like XGBoost achieved high AUC scores on ICU mortality prediction. These discoveries suggest that ML's capability to manage complex, large data may lead to more powerful predictions for the MCC=1 subgroup. Consequently, this study determines whether such achievents are certainly greater for patients with multiple chronic conditions.

3. Method

3.1. General Approach

3.1.1. Brief overview of your pipeline and objective

This study aims to predict ICU mortality using ML algorithms for two distinct groups of admissions of patients: MCC=0, including admissions of patients with one chronic condition along with possible non-chronic conditions, and MCC=1, including admissions of patients with more than one chronic condition only. The main goal is to find whether and which ML models can predict mortality more effectively for the MCC=1 group, which

generally has more complex health risks. As a consequence, four ML models were developed and assessed- LR, XGBoost, RF, and KNN. After comparing their performance in both groups individually, XGBoost outperformed the others in the key metrics of mortality prediction- precision, recall, and F1.

The steps in the pipeline are load the needed CSV files, remove rows with missing values, merge datasets by ‘subject_id’, ‘admission_id’ per patient, drop unnecessary columns, check for chronic and non-chronic conditions based on an additional CSV file, create new columns: MCC, ‘chronic_count’, ‘is_chronic’, generate ‘final_df’, train XGBoost on full dataset (‘final_df’) to compare the AUC to SAPS II AUC, split the full dataset (‘final_df’) into two groups df_mcc_0 and df_mcc_1, train/test LR, XGBoost, RF, KNN on ‘df_mcc_0’ and ‘df_mcc_1’, plot and compare model results, and identify best-performing model (XGBoost). Illustration can be seen in section 3.8.

3.1.2. MCC=0 VS MCC=1

To assess the ML models' performance better, the final dataset ready for testing was split into two subsets: MCC=0, including admissions of patients with exactly one chronic condition and possibly additional non-chronic conditions, and MCC=1, including admissions of patients with more than one chronic condition only. These groups were analyzed individually to see how each ML model performed within differing levels of chronic diseases.

For each group, LR, XGBoost, RF, and KNN were trained and assessed individually. This method approved the direct comparison of model performance within and between the MCC subgroups. Therefore, ML models are identified whether are more valuable for high-risk patients with MCC.

3.1.3. Predicting mortality using ML models

The target variable was binary in-hospital mortality throughout the analysis. A value

of 1 meant death in the ICU, and 0 indicated survival in the ICU. Each ML model was trained to categorize this outcome using the clinical features available in the dataset. Feature selection was applied to the XGBoost and RF models because these are adaptable to large feature spaces. However, LR and KNN were used without feature selection because of their poor performance in such settings.

Model evaluation was done using accuracy, AUC, precision, recall, and F1-score. Nonetheless, given the high clinical cost of false negative cases, greater importance was given to recall, precision, and F1-score to ensure that models could correctly find high-risk patients likely to experience ICU mortality.

3.2. Dataset Description

3.2.1. Name of dataset, size, source, time

The dataset used in this study is the Medical Information Mart for Intensive Care III Clinical Database (MIMIC III), version 1.4, which is a freely available dataset maintained by PhysioNet.org (Goldberger et al., 2000). It was collected from the Beth Israel Deaconess Medical Center in Boston, Massachusetts, and includes health-related data covering ICU admissions between 2001 and 2012 (Johnson, Pollard, & Mark, 2016).

Based on the data exploration by me, the dataset consists of 58,976 hospital admissions, 46,520 unique patients, and 61,532 ICU stays, along with a wide range of related clinical records. Based on the official documentation, there are 53,423 hospital admissions, 38,597 adult patients above 16 years old (Johnson et al., 2016). The article was published in 2016, but the dataset was modified in 2019, which is the reason why there are number differences. For this thesis, the newly updated numbers are used, updated on PhysioNet.org (Goldberger et al., 2000). MIMIC III is widely used in clinical research because of its broad coverage. It encompasses well-formed information on patient demographics, vital sign measurements recorded, laboratory test results, procedures,

medications, and mortality data, both in-hospital and after discharge. The information richness makes it especially valuable for model predictions, including ICU mortality prediction (Johnson, Pollard, & Mark, 2016).

3.2.2. General patient characteristics

The study focuses on adult ICU patients. Each admission record includes demographic, clinical, and diagnostic data collected during the ICU stay. The groups of admissions were categorized based on chronic disease count. MCC = 0 is the group of admissions involving patients with exactly one chronic condition, along with possible non-chronic conditions, and MCC = 1 is the group of admissions involving patients with more than one chronic condition only. This differentiation allows for a well-structured comparison of admissions of patient subgroups with different levels of disease complexity based on the features: ethnicity, first and last care unit, length of stay, mortality 0 or 1, count of chronic diseases, MCC- 1 for more than 1 chronic condition and 0 for 1 chronic condition along with possible other non-chronic conditions, count of non-chronic conditions, number of total conditions per admission, gender, mean vitals signs values, and mean laboratory values. The groups include 10660 admissions in the MCC=0 group and 48316 admissions in the MCC=1 group. From these features, the newly added are mortality, count of chronic diseases, MCC, count of non-chronic conditions, and number of total conditions per admission.

After making the data into the MCC=0 and MCC=1 groups, the same set of features was used to train and evaluate ML models for both subgroups. This consistent feature set guarantees correlation across models and groups when analyzing differences in predictive performance.

3.3. Data Cleaning and Preprocessing

3.3.1. Handling Missing Data

To ensure the consistency of the dataset, all rows that have missing values were

removed during the preprocessing phase. This cleaning step was done individually for each MIMIC III table used for this thesis before merging them. Only records with complete values for the selected clinical, demographic, and administrative features were kept for further analysis. This approach made the next processing and model training easier by preventing the need for fixing the data and making sure that each ML model was trained on fully checked data. This enhanced data completeness and assisted in avoiding possible bias introduced by inaccuracy or incompleteness of the dataset.

3.3.2. Feature Selection

Feature selection was applied to discover the most informative variables for ICU mortality prediction, to improve model outputs. For XGBoost and RF, feature selection was calculated using each model's built-in feature ranking mechanisms after training. These tools have numerical importance scores for each feature. Through these importance scores, all features for mortality prediction can be found, and only those with the lowest score are excluded. This helped avoid overfitting, especially in models that have the advantage of focusing on key variables. In addition, the feature importance analysis provided insights into which clinical and demographic factors had the strongest predictive value, with some features, such as mean vital signs, mean lab values, and chronic condition count, normally ranked supremely.

For LR and KNN, feature selection was not applied. These models were trained on the full set of input features due to their sensitivity to feature removal and do not usually support internal feature ranking. Using the complete feature set made sure of fairness and consistency in the input baseline before evaluation.

3.3.3. Any Transformations or Encodings

To prepare the dataset for model training, several data transformations and encodings were used. Categorical variables such as ethnicity, gender, first care unit, and last care unit

were encoded using one-hot encoding to convert them into a machine-readable format. Binary variables such as mortality, which is the target label, and the newly added feature MCC were already characterized as 0 and 1 and did not need further transformation.

Additionally, the synthetic minority over-sampling technique (SMOTE) was applied individually for each of the LR, XGBoost, RF, and KNN models to find class imbalance in the mortality outcome. This guaranteed that the models accepted balanced datasets during training and reduced bias toward the survivor class.

Feature selection was done only for the XGBoost and RF models based on their capability to internally rank feature importance. On the other hand, LR and KNN were trained on the full feature set because they are less adaptable with automatic feature elimination and can lose information.

These preprocessing steps guaranteed that all features were accordingly encoded, balanced, and structured to line up with the assumptions and capabilities of each ML model used in this research.

3.3.4. Explanation of MCC Grouping

Each hospital admission was assigned an MCC value (0 or 1) based on the number of chronic conditions present in the dataset. MCC=0 was used for admissions with exactly one chronic condition along with possibly non-chronic conditions, while MCC = 1 included more than one chronic condition only. This binary variable, derived from diagnostic codes, was used to split the dataset into two subgroups for separate model training and assessment.

3.4. Feature Selection in Details

3.4.1. How Feature Selection Was Chosen

The results from the XGBoost and RF models did not show expected accuracy, so it was necessary to enhance the results as much as possible. XGBoost and RF were trained using the same set of features to allow fair comparison between them and across the MCC=0

and MCC=1 groups. After training, feature importance tools were used to see which features had the biggest impact on the predictions. For XGBoost and Random Forest, '.feature_importances_' was used, which provided numeric scores for each feature. Features with importance scores lower than or equal to 0.01 were considered low-impact and were removed in the final versions of these models. This included several ethnic subgroups and care unit indicators that consistently ranked at the bottom. Some features that had a 0.01 score and were removed dropped the score of the model output, so they were manually returned.

For LR and KNN, all features were kep, since these models do not have built-in feature importance and can be sensitive to feature removal. This approach allowed me to compare all models under the same conditions while still learning which features mattered most.

3.4.2. Whether Selection Was Manual, Model-Based, or Both

Feature selection in this study was model-based. This combination of automated scoring and manual performance testing guaranteed that feature selection was both data-driven and optimized for each model's output.

3.5. Machine Learning Models

3.5.1. Used Models

Four ML models were used for the ICU mortality prediction: LR, XGBoost, RF, and KNN. LR and KNN were used as baseline models, while RF and XGBoost were selected for their strong performance with complex data, including many features. To compare results across different levels of chronic conditions, all ML models were applied to both MCC=0 and MCC=1 groups.

3.5.2. Reason To Select The ML Models

For balanced simplicity and performance, LR, XGboost, RF, and KNN models were

chosen. LR, XGBoost, RF, and KNN were applied by Alghatani et al. (2021), and LR, XGBoost, and RF by Iwase et al. (2022) for ICU mortality prediction. Therefore, it was guaranteed that these models are probably good to train on for mortality prediction in the context of chronic conditions.

3.5.3. Key Parameters

All four models were used with appropriate but light tuning to balance performance and comparability. As shown in table 1, for XGBoost, parameters like 'n_estimators'=50-150, 'max_depth'=6, and 'learning_rate'=0.05 were manually set. In addition, 'scale_pos_weight' was computed for MCC=1 to address class imbalance. 'n_estimators'=100 was used consistently in the subgroups for RF.

Classification thresholds were dynamically selected for XGboost and RF models in the MCC=1 group, aiming for recall ≥ 0.60 and precision ≥ 0.18 for RF, and recall ≥ 0.65 and precision ≥ 0.20 for XGBoost. If no such threshold was found, 0.4 was used as a backup.

LR was configured with 'max_iter'=1000 and 'class_weight'='balanced', and used a fixed threshold=0.4. KNN was used with 'n_neighbors'=5 and the same fixed threshold=0.4. All models included the SMOTE tool for class balancing and 'StandardScaler' for numerical features. Basic hyperparameter tuning was used. For example, for XGBoost, with parameters like the estimators numbers, learning rate, and tree depth were modified. In some cases, 'RandomizedSearchCV' was used to explore the most desirable configurations. This was useful for balancing performance while maintaining clarity in the MCC groups.

Table 1.

Parameters and Threshold

	XGBoost	RF	LR	KNN
Parameters	n_estimators=	n_estimators=	max_iter= 1000	n_neighbors= 5

	50-150 max_depth= 6 learning_rate= 0.05 scale_pos_weight	100	class_weight= 'balanced'	
Threshold	recall ≥ 0.65; precision ≥ 0.20	recall ≥ 0.60; precision ≥ 0.18	0.4	0.4

Note. The parameters and threshold fix per ML model used in this project code

3.6. Evaluation Metrics

3.6.1. Used metrics

Using accuracy, AUC, precision, recall, and F1, model performance was evaluated. Precision, recall, and F1 were prioritized because of the clinical importance of correctly identifying high-risk patients, while accuracy and AUC provide general performance insights. Precision, recall, and F1 are significantly relevant in imbalanced datasets such as ICU mortality because false negatives can have critical consequences. All four models were evaluated on MCC=0 and MCC=1 groups using the same metrics for consistency.

3.6.2. Prioritized metrics

The reasons to prefer precision, recall, and F1 over accuracy and AUC are that precision, recall, and F1 better represent the ability of the model to identify patients with high mortality risk. In the ICU environment, where a high-risk patient will be predicted to live, false negatives are fatal. High recall guarantees that most of the true mortality is captured. Precision limits false alarms. The F1 balances both, which renders these metrics more apt at evaluating model performance in this clinical scenario.

3.7. Implementation Details

3.7.1. Language: Python

The assessment, preprocessing, and data modelling were done in Python.

3.7.2. Used Libraries

Data manipulation was handled by using pandas, numpy, and dask libraries. Data preprocessing and modelling have been done using scikit-learn with modules train_test_split, StandardScaler, LogisticRegression, RandomForestClassifier, KNeighborsClassifier, and metrics classification_report, confusion_matrix, accuracy_score, roc_auc_score, precision_score, recall_score, and f1_score. Gradient boosting has been conducted using the xgboost library. Over-sampled or imbalanced data has been detected using the SMOTE method from the imblearn library. Hyperparameter tuning has been used using RandomizedSearchCV from scikit-learn. Bar plots were used to make visualizations with matplotlib and seaborn. All used libraries are illustrated in table 2.

Table 2.

Used libraries in this thesis code

Libraries
pandas
numpy
dask
scikit-learn
xgboost
imblearn
matplotlib
seaborn

Note. These are the libraries were used to import the necessary functions in Python.

3.7.3. Environment

The project was developed and executed on a local machine using Jupyter Notebook running on Windows 11. All scripts were written in Python 3.9.7, and the environment was managed using standard packages installed via pip.

3.7.4. Any cross-validation method

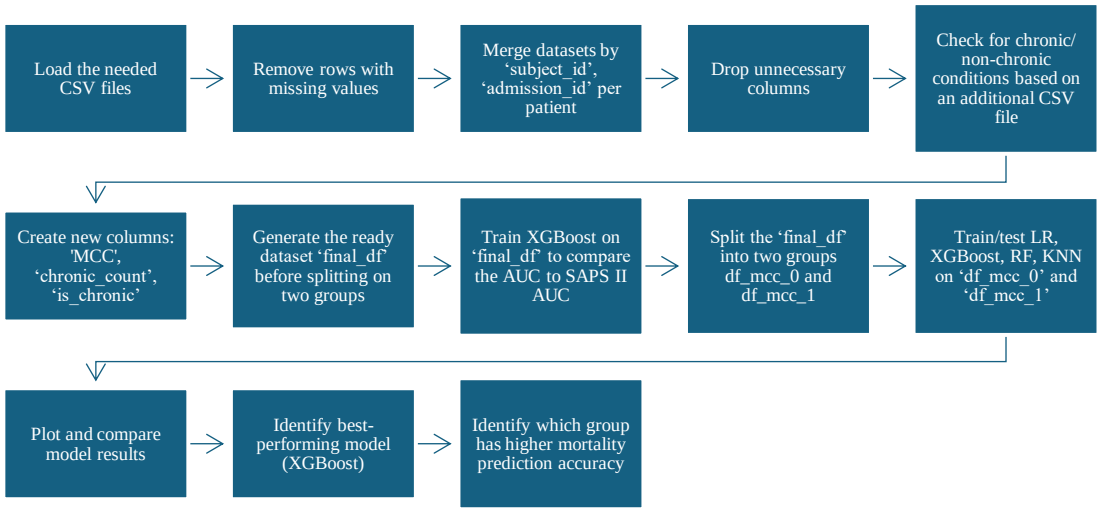
For model evaluation, a train-test split was used with an 80/20 split, making sure class balance between training and testing datasets. In addition, cross-validation was managed using 5-fold cross-validation by default during hyperparameter tuning, such as using RandomizedSearchCV for XGBoost. This helped in making a choice about strong model parameters and reducing the risk of overfitting.

3.7.5. Reproducibility, Random Seed

To guarantee reproducibility, a random seed of 42 was applied throughout the code . This included data splitting with train_test_split, SMOTE resampling, and model initialization where suitable. Handling consistent results across different runs of the ML pipeline was supported by using a fixed seed.

3.8. Workflow Diagram

Image 1. *Illustration of the project workflow*



Note. Workflow of the study design

4. Results

4.1. Results Overview

This section presents and interprets the performance of four ML models- LR and KNN as a baseline, RF, and XGBoost, used for ICU mortality prediction. The models were evaluated on two patient subgroups: MCC=0 admissions of patients with exactly one chronic condition, along with possible non-chronic conditions, and MCC=1 admissions of patients with multiple chronic conditions only.

Before the two groups' evaluation, XGBoost was trained on the dataset before splitting it on MCC=0 and MCC=1 to compare its AUC output with the traditional model SAPS II AUC result of 0.77 (Kong et al., 2020).

Each model was evaluated using the five standard classification metrics: accuracy, precision, recall, F1-score, and AUC. Performance is visualized using comparison bar plots. The focus is on precision, recall, and F1 metrics because of the mortality prediction.

4.2. Logistic Regression (LR) Performance (Baseline)

According to MCC=0, LR performed well in accuracy, recall, and AUC so it detected most mortality cases. Nevertheless, the model had a low precision and F1, showing many false positives. As a baseline, it gives a useful comparison point for the more advanced models. See image 2. See table 3.

According to MCC=1, precision, recall, and F1 increased. However, accuracy and AUC dropped. Despite a reasonable AUC, the model showed limited capability to make a difference between survival and death in this complex group. See image 3. See table 3.

Table 3.

Comparison between MCC=0 and MCC=1 according to LR

	MCC=0	MCC=1
Accuracy	0.85	0.50

Precision	0.09	0.12
Recall	0.58	0.74
F1	0.15	0.21
AUC	0.80	0.66

Note. Logistic regression results in both groups of patient admissions

4.3. XGBoost Performance

Before splitting the finalized 'final_df' table on MCC=0 and MCC=1, XGBoost model was performed on the 'final_df' to compare the AUC result with the SAPS II AUC result of 0.77 (Kong et al., 2020). Therefore, the XGBoost AUC result showed 0.81, which means that it is an ML algorithm outperforming one of the most accurate traditional methods. XGBoost was chosen for the comparison with the traditional model because XGBoost performed the best among the four tested models after splitting the dataset into groups.

According to MCC=0, XGBoost achieved a great accuracy, an AUC, and a recall. Nevertheless, precision and F1 appeared low, but both are higher than the precision and F1 of the baseline LR and KNN. See image 2. See table 4.

According to MCC=1, XGBoost's precision, recall, and F1 increased, demonstrating better sensitivity in identifying mortality among high-risk patients. Nonetheless, accuracy and AUC declined, consistent with the added complexity of this subgroup. Therefore, it is overall stronger than the baseline and RF and outperforms the MCC=0 in the most important metrics for mortality prediction. See image 3. See table 4.

Table 4.

Comparison between MCC=0 and MCC=1 according to XGBoost

	MCC=0	MCC=1
Accuracy	0.92	0.73

Precision	0.17	0.20
Recall	0.64	0.68
F1	0.27	0.31
AUC	0.90	0.79

Note. XGBoost results in both groups of patient admissions

4.4. Random Forest (RF) Performance

According to MCC=0, RF slightly outperformed XGBoost and the baseline, specifically in accuracy, precision, and F1. The high precision is stronger at limiting false positives than the baseline models. However, recall is lower than XGBoost and LR, but recall is one of the most important metrics for mortality prediction so RF missed many possible deaths. AUC is equal to the XGBoost result. See image 2. See table 5.

According to MCC=1, the performance decreased with increasing complex diseases. Accuracy, precision, F1, and AUC all declined below XGBoost. Only recall increased, but it remains under XGBoost. Therefore, it has a better balance than the baseline, but not as strong as XGBoost. See image 3. See table 5.

Table 5.

Comparison between MCC=0 and MCC=1 according to RF

	MCC=0	MCC=1
Accuracy	0.95	0.70
Precision	0.23	0.18
Recall	0.50	0.67
F1	0.32	0.28
AUC	0.90	0.76

Note. Random Forest results in both groups of patient admissions

4.5. k-Nearest Neighbors (KNN) Performance (Baseline)

According to MCC=0, KNN performed similarly to LR. Recall and AUC are lower than all other models. Accuracy, precision, and F1 are slightly higher than LR, lower than XGBoost, and RF. The AUC result demonstrates limited ability to separate the two outcome classes. Low performance on the three most important metrics for mortality means that KNN missed many possible deaths. See image 2. See table 6.

According to MCC=1, precision, recall, and F1 increased slightly. Accuracy and AUC decreased. Accuracy is much better than LR. KNN is similar to LR in precision, F1, and AUC, but much worse in recall than all models and overall worse than tree-based models like XGBoost and RF. See image 3. See table 6.

Table 6.

Comparison between MCC=0 and MCC=1 according to KNN

	MCC=0	MCC=1
Accuracy	0.91	0.67
Precision	0.12	0.14
Recall	0.44	0.50
F1	0.19	0.21
AUC	0.69	0.63

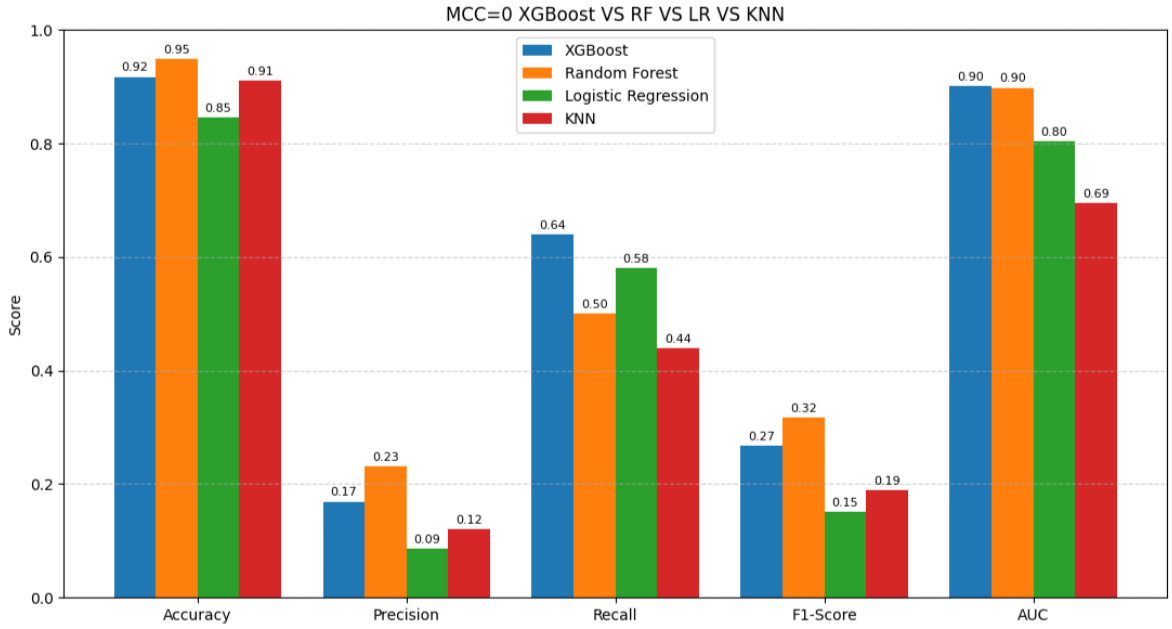
Note. k-Nearest Neighbors results in both groups of patient admissions

4.6. Model Comparison Across MCC Groups

All four ML models were tested on both MCC=0 and MCC=1 groups of patient admissions. While LR and KNN provided a useful baseline, they consistently underperformed compared to the tree-based models XGBoost and RF. See images 2 and 3.

Image 2.

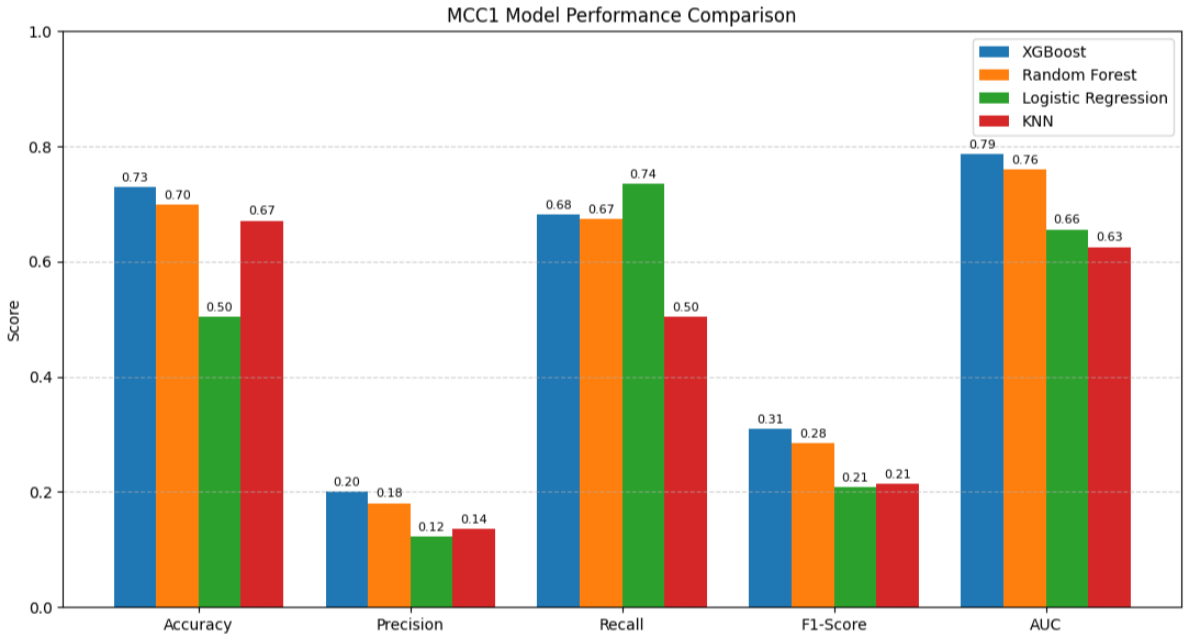
Visualization of all four ML models' results according to the MCC=0 group



Note. Comparison between XGBoost, RF, LR, and KNN according to the MCC=0 group

Image 3.

Visualization of all four ML models' results according to the MCC=1 group



Note. Comparison between XGBoost, RF, LR, and KNN according to the MCC=1 group

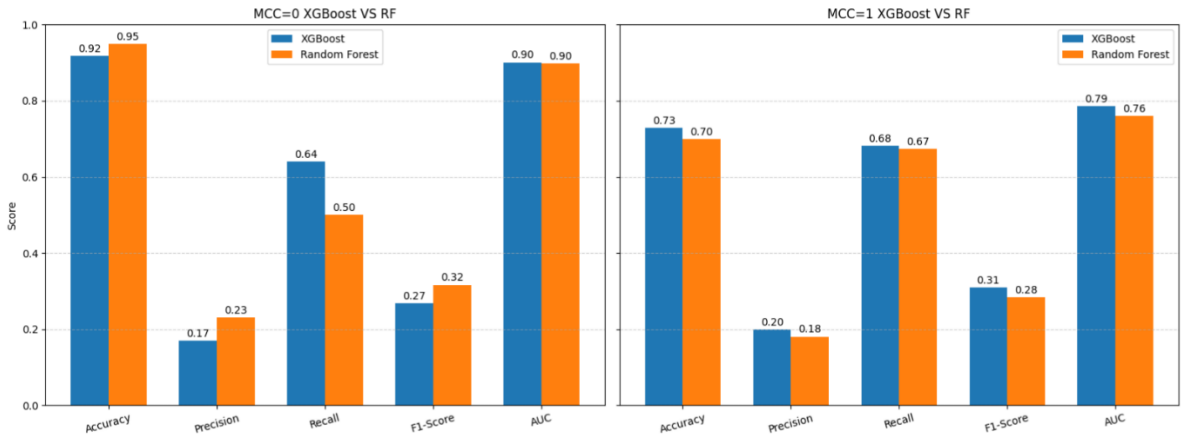
XGBoost and RF were selected for further comparison because of their superior results. Therefore, XGBoost showed itself as the most reliable, particularly in metrics critical for mortality prediction, such as recall, precision, and F1. Image 4 illustrates the

same results as in images 2 and 3, but excluding the LR and KNN models' performance.

For MCC=1 patients, XGBoost outperformed RF in all metrics. For MCC=0, XGBoost also maintained competitive performance with the recall score. However, AUC is equal to the RF result, and accuracy, precision, and F1 are lower than RF. But in mortality prediction, recall is one of the leading metrics that matter the most, and it can be concluded that XGBoost still outperforms RF in accurately predicting mortality. Because mortality prediction appeals for minimizing false negatives, precision, recall, and F1 were prioritized over accuracy and AUC. XGBoost is the most suitable model for predicting ICU mortality among all tested models in this thesis. See image 4.

Image 4.

Visualization of XGBoost and RF models' results according to both groups



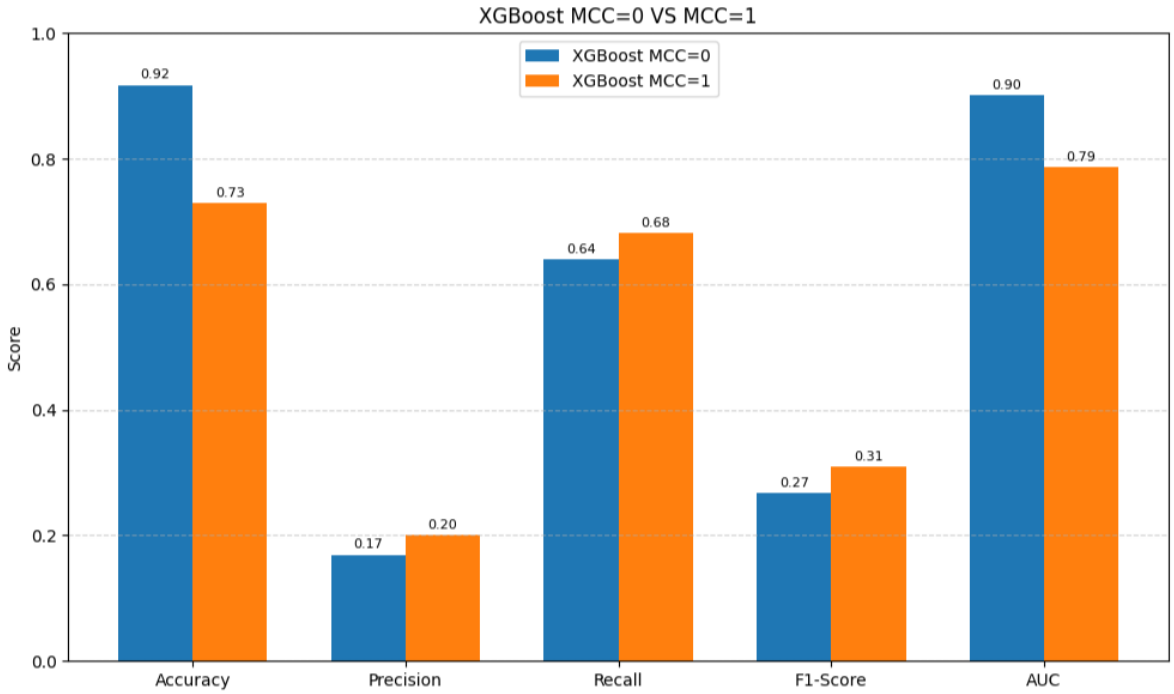
Note. Comparison between XGBoost and RF according to MCC=0 and MCC=1

4.7. MCC=0 and MCC=1 Comparison Based on XGBoost

The comparison between both groups in XGBoost confirms that the MCC=1 group has higher results than MCC=0 in the most important metrics- precision, recall, and F1. Therefore, this supports the alternative hypothesis (H_1): Machine learning models improve ICU mortality prediction accuracy for admissions of patients with multiple chronic conditions compared to admissions of patients with one chronic condition, along with possible non-chronic conditions. See image 5 and table 4.

Image 5.

Visualization of XGBoost results according to both groups



Note. Comparison between MCC=0 and MCC=1 in XGBoost model

4.8.Research Question

This study aims to answer whether ML models can accurately predict ICU mortality for admissions of patients with multiple chronic conditions (MCC=1) compared to admissions of patients with one chronic condition along with possible non-chronic conditions (MCC=0).

Among all models, XGBoost consistently achieved the best performance, particularly for the MCC=1 group. Precision, recall, and F1 metrics are particularly significant in predicting mortality because minimizing false negatives and discovering high-risk patients is more essential than overall accuracy or AUC. High recall guarantees fewer missed mortalities, high precision avoids false alarms, and a strong F1 balances the two.

On the other hand, LR and KNN showed worse results in the three key metrics than XGBoost and RF. LR and KNN are less suitable for discovering complex relationships in

clinical data. As a result, further comparison focused on XGBoost and RF, the two most capable models. Compared to RF, XGBoost performed better in the MCC=1 group across all metrics, and in MCC=0, it achieved notably higher recall than RF, which is essential for mortality prediction. Therefore, XGBoost is more efficient at discovering high-risk patients when more chronic conditions are present, supporting the alternative hypothesis (H_1): Machine learning models improve ICU mortality prediction accuracy for admissions of patients with multiple chronic conditions compared to admissions of patients with one chronic condition, along with possible non-chronic conditions. As a result, the answer to the research question is positive that ML models accurately predict ICU mortality for admissions of patients with multiple chronic conditions (MCC=1).

4.9.Post hoc analysis

Post hoc steps were taken for clarity enhancement. Feature importance analysis using XGBoost and RF showed that length of stay, mean vital signs, and chronic condition count were some of the most predictive features, particularly for the MCC=1 group. This was as a result of the models' internal scoring, for example, '.feature_importances_', not from prior assumptions, and aligned well with clinical expectations.

Additionally, classification thresholds for XGBoost and RF models were fine-tuned for better balance in recall, precision, and F1, especially essential for discovering mortality in high-risk MCC=1 admissions of patients. Thresholds were selected only when achieving the minimum criteria. For instance, for XGBoost, recall ≥ 0.65 and precision ≥ 0.20 were chosen. A default of 0.4 was used if no suitable threshold was found. These post hoc modifications enhanced practical performance without changing the overall pipeline, strengthening the models' clinical service.

5. Discussion

5.1. Study Objective

This study had the goal to answer the research question, ‘Can machine learning accurately predict ICU mortality in admissions of patients with multiple chronic conditions compared to admissions of patients with one chronic condition, along with possible non-chronic conditions?’. Admissions of patients with multiple chronic conditions are in the MCC=1 group, and admissions of patients with one chronic condition along with possible non-chronic conditions are in the MCC=0 group. Therefore, it had to be discovered which hypothesis the research can support. Either the H_0 (Null Hypothesis): Machine learning models do not improve ICU mortality prediction accuracy for admissions of patients with multiple chronic conditions compared to admissions of patients with one chronic condition, along with possible non-chronic conditions, or H_1 (Alternative Hypothesis): Machine learning models improve ICU mortality prediction accuracy for admissions of patients with multiple chronic conditions compared to admissions of patients with one chronic condition, along with possible non-chronic conditions.

5.2. Key Findings

For MCC=1, with the highest precision, recall, and F1, XGBoost performed best compared to the other three ML models. The results support the alternative hypothesis (H_1). This validates that ML models, especially XGBoost, are more efficient for predicting the mortality of high-risk patients with MCC.

5.2.1. Comparison to Previous Studies

The performance of the XGBoost model in this study is aligned with previous ICU mortality prediction studies. This is because in other studies there is an AUC of 0.87, which is one of the best performances (Ye et al., 2020), a recall of 0.80, which is good and satisfying (Darabi et al., 2018), and F1 was found up to 0.85, which is a good result and much better than other ML models (Subudhi et al.,2021). So, the recall and F1 in the current research can improve slightly more, but they are in the right direction. The low precision and

F1 in this thesis are expected in imbalanced datasets, where high AUC may still miss true deaths (Subudhi et al., 2021; Nistal-Nuño, 2022). Reduced variance and affected calibration might be caused by dividing the dataset into groups based on chronic conditions, especially in more complex or smaller patient groups. See table 8.

Table 8.

Comparison between XGBoost model range of results of the MCC groups and other studies results

	XGBoost	Other studies
AUC	0.79-90	0.87
Recall	0.64-0.68	0.80
F1	0.27-0.31	Up to 0.85

Note. MCC=1 and MCC=0 are combined in the column 'XGBoost' range and compared with other studies results in mortality prediction with XGBoost

5.3. Surprising Patterns

Although recall is high, precision and F1 remained low overall in all models. This may be due to the data imbalance between the two groups of patient admissions, despite applying the SMOTE method for enhancing the outputs.

5.4. Limitations

This study has several limitations. First, while XGBoost achieved better precision, recall, and F1-score for MCC=1 patients, which are the most essential metrics in mortality prediction, it still performed slightly better on accuracy and AUC for MCC=0 when compared with the MCC=1 group. Ideally, the model would perform stronger than MCC=0 across all metrics for MCC=1, since that is the focused group.

Second, the analysis relied only on structured, tabular data and excluded more complex sources like time-series trends or clinical notes, which may include useful context

for mortality prediction.

Third, this thesis used data from only one hospital, which could limit how well the results generalize to other healthcare systems.

Fourthly, simpler models like LR and KNN underperformed, which shows that the model choice significantly affects outcomes and highlights the need for better handling of complex feature interactions in future work.

Lastly, splitting patient admissions into MCC=0 and MCC=1 has increased class imbalance and may have reduced within-group variability, which could partly explain the lower precision and F1 metrics compared to previous studies. Even though there are several limitations, the study answered the research question and provided strong evidence in support of the alternative hypothesis (H_1).

5.5. Contribution to Literature

Different from previous studies, in this thesis the the ICU patients were split into two subgroups and their performance was analyzed individually for each ML model. This problem of the thesis supports the precision, recall, and F1 metrics as more essential metrics than just AUC or accuracy in mortality prediction. The findings demonstrate that assessing ML models by subgroup gives a more relevant understanding of ML performance in complex clinical settings. As a result, this helped to show that XGBoost is notably effective for more complex cases like the MCC=1 group of admissions of patients. This would not have been understandable if only overall performance was considered and studied.

5.6. Future Work

Future work should apply the same model pipeline to external ICU datasets to test generalizability. Including time-series data like vital signs over time and clinical notes could assist in discovering more informative patient details. Additionally, using model explanation tools like SHAP would help explain how individual features contribute to each prediction,

increasing clinical transparency and trust in the model.

Future studies should also experiment with different subgrouping strategies, different from the chronic condition count. For instance, clustering by profiles with complex diseases. Therefore, the model accuracy could be enhanced, and it might help handle class imbalance more efficiently, particularly in smaller or higher-risk subgroups.

6. Conclusion

Overall, in the thesis, the alternative hypothesis (H_1) answered the research question. By training and evaluating four ML models, Logistic Regression, k-Nearest Neighbors, Random Forest, and XGBoost, on two subgroups: admissions of patients with one chronic condition along with possible non-chronic conditions, and admissions of patients with more than one chronic condition only. The findings demonstrated that XGBoost consistently performed best compared to the other three models, especially for the group of admissions of patients with more than one chronic condition. Precision, recall, and F1 metrics were prioritized in mortality prediction to avoid missing possible deaths.

This thesis contributes to the literature by introducing subgroup-based evaluation and strengthening the value of more nuanced metrics in high-risk settings. Future research should have the goal to generalize findings using external datasets, more detailed and informative data, and interpretable ML techniques to improve trust and clinical adoption.

My thesis achieved its goal by demonstrating that machine learning, particularly XGBoost, can support ICU doctors by more accurately identifying patients in the high-risk group, eventually contributing to earlier interventions and enhanced patient outcomes.

References

Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). *PhysioNet*. https://doi.org/10.13026/C2XW26

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*(1), 160035. https://doi.org/10.1038/sdata.2016.35

Kong, G., Lin, K., & Hu, Y. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. BMC Medical Informatics and Decision Making, 20(1), 251. https://doi.org/10.1186/s12911-020-01271-2

Nistal-Nuño, B. (2022). Developing machine learning models for prediction of mortality in the medical intensive care unit. Computer Methods and Programs in Biomedicine, 216, 106663. https://doi.org/10.1016/j.cmpb.2022.106663

Gao, J., Lu, Y., Ashrafi, N., Domingo, I., Alaei, K., & Pishgar, M. (2024). Prediction of sepsis mortality in ICU patients using machine learning methods. BMC Medical Informatics and Decision Making, 24(1), 228. https://doi.org/10.1186/s12911-024-02630-z

Mamandipoor, B., Frutos-Vivar, F., Peñuelas, O., Rezar, R., Raymondos, K., Muriel, A., Du, B., Thille, A. W., Ríos, F., González, M., del-Sorbo, L., del Carmen Marín, M., Pinheiro, B. V., Soares, M. A., Nin, N., Maggiore, S. M., Bersten, A., Kelm, M., Bruno, R. R., … Osmani, V. (2021). Machine learning predicts mortality based on analysis of ventilation parameters of critically ill patients: Multi-centre validation. BMC Medical

Informatics and Decision Making, 21(1), 152. https://doi.org/10.1186/s12911-021-01506-w

Alghatani, K., Ammar, N., Rezgui, A., & Shaban-Nejad, A. (2021). Predicting Intensive Care Unit Length of Stay and Mortality Using Patient Vital Signs: Machine Learning Model Development and Validation. JMIR Medical Informatics, 9(5), e21347. https://doi.org/10.2196/21347

Subudhi, S., Verma, A., Patel, A. B., Hardin, C. C., Khandekar, M. J., Lee, H., McEvoy, D., Stylianopoulos, T., Munn, L. L., Dutta, S., & Jain, R. K. (2021). Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. Npj Digital Medicine, 4(1), 1–7. https://doi.org/10.1038/s41746-021-00456-x

Darabi, H. R., Tsinis, D., Zecchini, K., Whitcomb, W. F., & Liss, A. (2018). Forecasting Mortality Risk for Patients Admitted to Intensive Care Units Using Machine Learning. Procedia Computer Science, 140, 306–313. https://doi.org/10.1016/j.procs.2018.10.313

Iwase, S., Nakada, T., Shimada, T., Oami, T., Shimazui, T., Takahashi, N., Yamabe, J., Yamao, Y., & Kawakami, E. (2022). Prediction algorithm for ICU mortality and length of stay using machine learning. Scientific Reports, 12(1), 12912. https://doi.org/10.1038/s41598-022- 17091-5

Ye, J., Yao, L., Shen, J., Janarthanam, R., & Luo, Y. (2020). Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. BMC Medical Informatics and Decision Making, 20(11), 295. https://doi.org/10.1186/s12911-020-01318-4