

**MORTALITY PREDICTION MODEL FOR GENERAL SURGERIES**  
**USING A SMALL DATA SET WITH EXPLAINABLE**  
**ARTIFICIAL INTELLIGENCE**

by

Anil Kumar Pandey, MBBS, MD, DNB, MBA, PGDFM, PGCPF, PGCPIB,  
PGCPMLDL

DISSERTATION

Presented to the Swiss School of Business and Management Geneva

In Partial Fulfillment

Of the Requirements

For the Degree

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

November, 2024

**A MORTALITY PREDICTION MODEL FOR GENERAL SURGERIES  
USING A SMALL DATA SET WITH EXPLAINABLE ARTIFICIAL  
INTELLIGENCE**

by

Anil Kumar Pandey

Supervised by

Dr. Kamal Malik

APPROVED BY

dr. Ljijana Kukec, Ph.D.



Dissertation chair

RECEIVED/APPROVED BY:

---

Admissions Director

## **Dedication**

This dissertation is dedicated to all those who have inspired and supported me throughout this academic achievement.

To my family, for their unwavering love, patience, and encouragement; to my mentors and colleagues, for their guidance and wisdom; and to my friends, for their constant support and understanding.

This work would not have been possible without each of you. Thank you for believing in me and for being my pillars of strength.

## **Acknowledgments**

I am deeply thankful to all those who have contributed to completing this dissertation.

First and foremost, I would like to express my heartfelt gratitude to my Mentor, Dr. Kamal Malik, for their invaluable guidance, support, and encouragement throughout this journey. Your insights and wisdom have been instrumental in shaping this work, and your belief in my abilities has been a constant source of motivation.

To my parents for your endless love and support, my spouse for your unwavering partnership, and to my children for the boundless joy you bring. Your sacrifices and encouragement have been the foundations of my success.

Finally, I thank my colleagues and friends for their support and encouragement. Your camaraderie has made this journey both enriching and enjoyable.

I want to express my heartiest gratitude to everyone who has supported me throughout my journey.

## **ABSTRACT**

# **A MORTALITY PREDICTION MODEL FOR GENERAL SURGERIES USING A SMALL DATA SET WITH EXPLAINABLE ARTIFICIAL INTELLIGENCE**

Anil Kumar Pandey  
2024

Dissertation Chair: <Chair's Name  
Co-Chair: <If applicable. Co-Chair's Name>

This study introduces a mortality prediction model for general surgeries, enhanced by Explainable Artificial Intelligence (XAI) to help surgeons anticipate postoperative outcomes and identify high-risk patients. Using a deep learning approach on a limited dataset, synthetic data generation via a variational autoencoder (VAE) was employed to simulate real-world accuracy. The deep learning model trained with VAE-augmented data emerged as the best performer in comparison to other machine learning models, including RF, KNN, extreme gradient boosting, support vector machines, and logistic regression., achieving the highest F1 score and balanced precision and recall.

Patient data—including morbidities, laboratory results, and postoperative complications—was processed through various models and evaluated on accuracy, F1 score, and AUROC metrics. The VAE data augmentation improved the performance of most models, especially complex ones such as decision trees, random forests, gradient boosting, and XGBoost. However, simpler models like logistic regression and support vector machines (SVC)

struggled with VAE-augmented data. The ensembling approach incorporating Ensembles of VAE, Flipout in last layer, Flipout in all layers and Bayesian model was used to improve prediction, The ensemble model, *VAE-Flipout Last Layer and Flipout All Layers Ensemble*, demonstrated enhanced predictive accuracy and reliability surpassing other ensemble models . Calibration techniques, including Temperature Scaling, Platt Scaling, and Isotonic Regression, were applied to ensure robust probabilistic outputs. The *VAE-Flipout Last Layer and Flipout All Layers Ensemble* achieved an F1 score of 0.77, a Brier score of 0.0254, and a ROC-AUC of 0.94.

In terms of model explainability, LIME and SHAP identified the features influencing mortality, including Sepsis, Postoperative Urea, Small Bowel Resection, Omentoplasty ASA Classification, Chronic Liver Disease, and postoperative biomarkers like SGPT and bilirubin. LIME provided local insights tailored to individual predictions, while SHAP revealed a global perspective across all instances, consistently highlighting these key features. This consistency reinforces the relevance of identified factors in patient outcomes. Future research should focus on expanding the dataset through advanced augmentation techniques, like GANs, and on refining calibration metrics, such as Expected Calibration Error (ECE) and Maximum Calibration Error (MCE). This study offers a valuable AI-driven tool to improve patient prognosis and postoperative outcomes.

## TABLE OF CONTENTS

List of Tables	x
List of Figures	xi
<b>CHAPTER I: INTRODUCTION</b>	<b>1</b>
<b>1.1 Introduction</b>	<b>1</b>
1.1.1 Challenges in current approach	4
1.1.2 Overfitting	6
1.1.3 Uncertainty Quantification	7
1.1.4 Computational Expense	8
1.1.5 Generalization Across Domains	9
<b>1.2 Research Problem</b>	<b>10</b>
<b>1.3 Purpose of Research</b>	<b>13</b>
1.4 Significance of the Study	13
<b>1.5 Research Purpose and Questions</b>	<b>14</b>
<b>CHAPTER II: REVIEW OF LITERATURE</b>	<b>15</b>
<b>2.1 Introduction &amp; Theoretical Framework</b>	<b>19</b>
2.1.1 Machine Learning Techniques	19
2.1.1.1 Tree-based Models: Decision Tree	31
2.1.1.2 Random Forest	32
2.1.1.3 Boosting Algorithm	36
2.1.1.4 Support Vector Machine	39
2.1.1.5 Kernels	42
2.1.1.6 Deep Learning Models	45
2.1.1.7 The challenge of over fitting in SVMs & Deep Learning	59
2.1.1.8 Applications of SVMs, Deep Learning and BNNs	60
2.1.2 Explainable AI Methods: Overview	61
2.1.3 Model- Agnostic Interpretability of Machine Learning	65
2.1.3.1 LIME	66
2.1.3.2 SHAP	68
2.1.3.3 SHAP Values	69
<b>2.2 Summary</b>	<b>73</b>
<b>CHAPTER III: METHODOLOGY</b>	<b>79</b>
3.1 Overview of the Research Problem	79
3.2 Operationalization of Theoretical Constructs	81
3.2.1 Feature Importance	83
3.2.2 Additional Statistics	84



3.3 Research Purpose and Questions .....	84
3.4 Research Design.....	85
3.4.1 Population .....	85
3.4.2. Data Sources.....	85
3.4.3 Data Challenges .....	86
3.4.4 Limitations of the dataset size and balance.....	85
3.4.5 Model Development Machine Learning Techniques.....	86
3.4.6 Model Evaluation Metrics .....	86
3.4.7. Calibration Techniques .....	87
3.5 Population and Study Sample .....	87
3.5.1 Sample Size and Selection of Sample.....	87
3.6 Participant Selection .....	88
3.7 Instrumentation .....	89
3.8 Data Collection Procedure .....	89
3.9 Data Analysis .....	90
3.9.1Data Analysis Strategies .....	91
3.10 Research Design Limitations.....	92
<b>CHAPTER IV: RESULTS.....</b>	<b>93</b>
4.1 Research Question One.....	97
4.2 <b>Research Question Two</b> .....	100
4.3 Research Question Three .....	135
4.4 Summary of findings .....	158
4.5 Conclusion .....	166
<b>CHAPTER V: DISCUSSION .....</b>	<b>168</b>
5.1 Discussion of Results .....	168
5.2 Discussion of Research Question One.....	170
5.2 Discussion of Research Question Two .....	172
5.3 Discussion of Research Question Three.....	177
<b>CHAPTER VI: SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS 181</b>	
6.1 Summary .....	181
6.2 Implications.....	183
6.3 Recommendations for Future Research.....	185
6.4 Conclusion .....	186
<b>APPENDIX A STATA &amp; R ANALYSIS.....</b>	<b>190</b>
<b>APPENDIX B INFORMED CONSENT.....</b>	<b>191</b>
<b>APPENDIX C MODEL CODE .....</b>	<b>193</b>

**APPENDIX D MODEL INTERPRETABILITY USING LIME AND SHAP  
ANALYSIS ..... 193**

**REFERENCES [USE “CHAPTER TITLE” STYLE] ..... 194**

## LIST OF TABLES

Table Number	Description	Page
Table 4.1	Performance of Machine learning models without correction of class Imbalance (Original data)	99
Table 4.2	Comparison of Machine learning model with Deep Learning Model with data imbalance corrected from Data Augmentation by Variational Autoencoder (VAE)	100
Table 4.3	Data augmentation techniques and performance of DNN models	102
Table 4.4	Comparison of Probabilistic Models	108
Table 4.5	Comparison of ensemble models	114-115
Table 4.6	Evaluation of models with as per matrices with Temp Scaling, Platt Scaling, and Isotonic regression	118-120
Table 4.7	Calibration of ensemble models with Temperature Scaling and brier score before and after	121
Table 4.8	Brier Score (Before and After Isotonic Calibration)	123
Table 4.9	Lime Output With Values Of Coefficients	142-145
Table 4.10	Sorted Positive SHAP values for variables	147-148

## LIST OF FIGURES

Figure Number	Description	Page number
Figure 2.1	SHAP (Shapley Additive explanation Values	70
Figure 4.1	Comparison of SMOTE and Its Variants	105-106
Figure 4.2	Comparison of ROC of VAE and Bayesian Models	112
Figure 4.3	Brier Score for Different Ensemble Models before and after Temp Scaling	122
Figure 4.4	Brier Score for Different Ensemble Models before and after Temp Scaling	124
Figure 4.5	Brier Score for Different Ensemble Models before and after Isotonic regression	124
Figure 4.6	Brier Score for Different Ensemble Models before and after Isotonic regression- Zoomed Calibration	125
Figure 4.7	ROC_AUC Ensemble of VAE Model, Model_1 and Model_2	127
Figure 4.8	Comparison of ROC of Ensemble Models I	131
Figure 4.9	Comparison of ROC of Ensemble Models II	132
Figure 4.10	Comparison of ROC of Ensemble Models III	133
Figure 4.11	Data instance 10th row with interpretation from LIME	137
Figure 4.12	Data Instance 10th with interpretation from SHAP	138

Figure 4.13	SHAP Values of Important Variables	138
Figure 4.14	Data Instance 6th with interpretation from LIME	139
Figure 4.15	Data Instance 6th with interpretation from SHAP	140
Figure 4.16	Data Instance 6th with interpretation from SHAP Showing SHAP Values in Waterfall	140
Figure 4.17	SHAP VALUES OF VARIABLES	157

## CHAPTER I:

### INTRODUCTION

#### 1.1 INTRODUCTION

A small subset of high-risk patients accounts for the majority of surgical complications (Pearse et al., 2006). Research has shown that timely interventions can significantly reduce or even prevent perioperative complications (Kang et al., 2012; Leeds et al., 2017). Therefore, it is essential to develop methods that can quickly identify patients who are most at risk for perioperative complications (Hill et al., 2019). Non-Cardiac General surgeries are the most common surgical procedures. Unlike manufacturing where production and customer information are dealt with by separate divisions in the organization. In the healthcare sector, it is the treating surgeon who is responsible for both the actual conduct of surgery and patient information including prognostication of the outcome of surgery. Patients present with the main problem considered a diagnosis and often have other morbid conditions. These comorbidities have important implications for the outcome of surgeries. Baseline comorbidity adjustment plays a crucial role in both health services research and clinical prognosis (Austin et al., 2015). Mortality prediction models that incorporate comorbidities from patient profiles can assist surgeons in effectively communicating potential outcomes to patients and their families. The Charlson and Elixhauser comorbidity classification systems are among the most widely used in health research, with studies showing their significant association with various outcomes, including in-hospital mortality (Poses et al., 1996; Sundararajan et al., 2004). Traditionally, statistical approaches have been employed for outcome prediction; however, in recent years,

machine-learning techniques have become increasingly popular in clinical applications (Guillaume et al., 2017; Chen et al., 2015; Mišić et al., 2020; Frizzell et al., 2017). Further, with the wider availability of computing power, there is a trend towards making use of DNN models.

In a somewhat simplified manner, machine learning tends to prioritize predictive accuracy over hypothesis-driven inference, often applying its methods to large, high-dimensional datasets with numerous covariates (Williams and Rasmussen, 2006; Breiman, 2001). This shift in focus has been critical in fields such as genomics, image processing, and, more recently, healthcare. Regarding surgical outcomes, current risk prediction models include those developed by the American Society of Anesthesiologists (American Society of Anesthesiologists, 1963) and the Surgical Apgar Score. Additionally, the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) has introduced its own Surgical Risk Calculator (ACS-SRC) (Bilimoria, 2013). Despite the utility of these models, many rely heavily on traditional statistical approaches that may fail to capture complex, non-linear interactions between variables (Shilo et al., 2020). Most of these models are grounded in traditional statistical methods. However, machine learning and deep learning, both branches of artificial intelligence, have seen significant uptake in various fields in recent years. For example, Lee et al. (2018) developed a deep neural network model trained on intraoperative data to predict postoperative in-hospital mortality. This model, however, uses complex variables that may only be practical in advanced surgical settings. While deep neural networks are widely used across industries, their adoption in medicine has been slower. Some researchers criticize machine learning as a

“black box,” questioning its applicability in clinical practice (Cabitza et al., 2017). Recent advancements in artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), have sparked interest in applying these techniques to improve clinical outcomes. ML and DL models have shown remarkable accuracy in predicting postoperative complications, with their ability to analyse large amounts of data and detect patterns that conventional statistical models might overlook (Rajkomar et al., 2018; Topol, 2019). Lee et al. (2018) developed a deep neural network model trained on intraoperative features to predict postoperative in-hospital mortality, demonstrating the potential of DL to significantly improve predictive accuracy. However, the model requires complex variables that may be accessible only in advanced surgical centers, limiting its generalizability.

In clinical settings, the adoption of AI has been slower compared to other industries, in part due to concerns about the interpretability of these models. While the predictive power of machine learning is impressive, many view these models as “black boxes” that lack transparency, raising doubts about their clinical utility and acceptance (Cabitza et al., 2017). To address this, new techniques such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) have been introduced, offering a way to make machine learning models more interpretable by identifying the importance of individual features in the prediction process (Lundberg & Lee, 2017; Ribeiro et al., 2016). These techniques enable clinicians to better understand the factors driving predictions, fostering trust, and enabling integration of ML models in healthcare decision-making.



As deep learning continues to evolve, the potential for developing more robust and interpretable models for predicting postoperative outcomes grows. These models offer opportunities not only for improving clinical decisions but also for identifying key risk factors that can guide interventions to reduce perioperative complications, thus improving patient outcomes (Miotto et al., 2018). However, for a model to have any use for professionals like surgeons, the need for an explanation of the model in the prediction of surgical outcomes is very high. Recent advancements in model explainability techniques can meet this objective.

### **1.1.1 Challenges in Current Approaches**

When addressing complex problems in machine learning, especially in fields like healthcare or financial forecasting, several significant challenges impede achieving optimal results. These challenges include data scarcity, overfitting, uncertainty quantification, and computational expense. Each of these issues presents both theoretical and practical difficulties that must be overcome for models to be both effective and reliable in real-world applications.

#### **A. Data Scarcity:**

##### **i. Theoretical Perspective:**

Machine learning models, intense learning approaches, typically require large amounts of labelled data to generalize well. The absence of sufficient data as in this study can lead to poor model performance since the models cannot learn the underlying distribution of the data. This is especially true in domains like hospitalized patient care, where annotated data

(e.g., patient records, diagnostic images) is limited due to privacy concerns, the high cost of obtaining data, and ethical considerations (Shorten & Khoshgoftaar, 2019). With small datasets, models are prone to high variance, leading to significant overfitting. Data scarcity also hinders the model's ability to identify rare patterns, potentially missing out on valuable signals in the data.

**ii. Practical Perspective:**

Collecting a large and diverse dataset may not always be feasible. For example, in medical diagnosis, certain conditions (like rare diseases) affect a small fraction of the population, making it difficult to gather enough examples for robust training (LeCun, Bengio & Hinton, 2015). The volume of data is a particular constraint which, however, entirely depends on the workload of the hospital. Data augmentation techniques, such as Variational Autoencoders (VAE), Generative Adversarial Networks (GANs), and Synthetic Minority Over-sampling Technique (SMOTE), are employed to mitigate data scarcity. However, these methods have limitations, including the risk of generating synthetic samples that may not reflect the true distribution of the target population (Chawla et al., 2002; Goodfellow et al., 2014). Additionally, external data sources (such as publicly available medical datasets) may not always match the local data context, leading to potential biases or inaccuracies when models are deployed in practice (He et al., 2020).

### **1.1.2 Overfitting**

#### **i. Theoretical Perspective:**

Overfitting occurs when a model learns the noise in the training data rather than the underlying signal, resulting in excellent performance on the training set but poor generalization to new, unseen data. This is particularly prevalent in deep learning models due to their large number of parameters (Srivastava et al., 2014).

In small datasets, overfitting becomes even more severe because the model has fewer examples to learn from, which can lead to exaggerated differences between the training and testing performance (Hawkins, 2004).

#### **ii. Practical Perspective:**

Overfitting is a concern when models are deployed in high-stakes environments like healthcare, where false positives or negatives can have critical implications (Zhou et al., 2020). For instance, overfitting can result in unreliable predictions of patient mortality, leading to incorrect clinical decisions.

Regularization techniques such as dropout, L2 regularization, and early stopping are used to combat overfitting, but they may not always work well with highly complex models (Goodfellow, Bengio & Courville, 2016). Furthermore, extensive hyperparameter tuning is often required, increasing time and computational demands (Bengio, 2012).

### **1.1.3. Uncertainty Quantification**

#### **i. Theoretical Perspective:**

In deep learning, especially in applications like healthcare and autonomous systems, uncertainty quantification plays a critical role. Traditional deep learning models, including neural networks, often produce highly confident predictions, even when they are incorrect, which poses a significant risk in decision-making processes that demand caution (Gal & Ghahramani, 2016). This is particularly dangerous in high-stakes situations, where inaccurate yet confident predictions can lead to adverse outcomes.

Machine learning, including deep learning, shares a close relationship with statistics in its goal to construct models that capture the processes generating observed data. Probability theory plays an essential role here, allowing us to design models that best fit the data by using probabilistic rules. While probability allows us to model processes by capturing uncertainties as random variables, statistics focuses on observing outcomes and inferring the underlying processes that explain these observations.

In machine learning, a key objective is to minimize generalization error, which refers to how well a model performs on new, unseen data rather than just the training data. This makes uncertainty quantification vital, as it helps ensure that the model's predictions are reliable when applied to future scenarios. Deep learning models typically lack a built-in mechanism to quantify uncertainty, leading to overconfident results in situations where predictions might be uncertain.

Bayesian methods, particularly Bayesian Neural Networks (BNNs), address this issue by treating model parameters as distributions rather than fixed values. This approach enables the estimation of predictive uncertainty, providing a more robust framework for managing uncertainty in high-risk applications (Blundell et al., 2015). Such probabilistic models can help reduce overconfidence and allow for safer, more cautious decision-making in scenarios where the stakes are high.

**iii. Practical Perspective:**

Implementing Bayesian models or probabilistic techniques is challenging due to the computational expense. Models like Monte Carlo Dropout or Flipout approximate Bayesian inference but introduce trade-offs between accuracy and speed (Gal & Ghahramani, 2016; Wen et al., 2018).

Despite their potential, many deep learning applications neglect uncertainty quantifications, leading to overconfident, high-risk decisions (Amodei et al., 2016). Calibration techniques, such as Temperature Scaling and Platt Scaling, are essential for addressing this issue but require additional computational resources (Guo et al., 2017).

### **1.1.4 Computational Expense**

**i. Theoretical Perspective:**

Deep learning models, particularly large architectures like ResNet and DenseNet, are computationally expensive to train. As models grow in complexity, so do the training time and resource requirements, making it difficult to deploy these models in real-time applications (He et al., 2016).

Techniques like variational inference and ensemble learning often introduce additional computational overhead, further complicating deployment in practical settings (Dietterich, 2000).

**iii. Practical Perspective:**

The computational burden is often overwhelming for smaller research labs or companies lacking access to high-performance hardware (Silver et al., 2016). Cloud-based services, while helpful, add significant financial costs.

Real-time applications, such as automated diagnostics, require quick inference times, which probabilistic and ensemble models may struggle to provide (Lakshminarayanan, Pritzel & Blundell, 2017).

**1.1.5. Generalization Across Domains**

**i. Theoretical Perspective:**

Generalization remains a critical challenge. A model trained on one dataset may not perform well on another due to differences in data distributions (Daumé, 2009). Domain adaptation techniques, while promising, introduce additional complexity and uncertainty (Pan & Yang, 2010).

**ii. Practical Perspective:**

Transfer learning and domain adaptation techniques aim to help models generalize to different domains, but in low-data scenarios, their effectiveness is limited. This is especially true in healthcare, where variations in patient demographics and diagnostic practices lead to significant performance drops (Goodfellow, Bengio & Courville, 2016).

In short, the challenges of data scarcity, overfitting, uncertainty quantification, computational expense, and generalization across domains are central to the difficulties encountered in modern machine learning. Addressing these challenges requires a multi-faceted approach, combining innovative model architectures, data augmentation techniques, regularization strategies, and probabilistic modeling (Zhou et al., 2020). However, the complexity of the solutions often introduces new difficulties, such as increased computational demands and the need for rigorous uncertainty quantification, which must be carefully balanced to ensure models are reliable and efficient in practical applications.

## **1.2 Research Problem**

Artificial Techniques (AI) including Deep learning have been tried in mortality prediction with equivocal results. In recent years, there has been a growing interest in leveraging artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), to predict patient outcomes and improve clinical decision-making. Ahmed et al. (2020) demonstrated the effectiveness of deep neural networks in predicting mortality among trauma patients admitted to intensive care units. Their model, which was supplied with statistically significant risk factors derived from patient data, outperformed conventional machine learning techniques such as Linear Discriminant Analysis (LDA), Gaussian Naive Bayes (GNB), K-Nearest Neighbors (KNN), and Decision Trees (CART). The deep neural network achieved a training accuracy of 93.8% and testing accuracy of 92.3%, with a sensitivity of 79.1%, specificity of 94.2%, positive predictive value (PPV) of 66.42%, negative predictive value (NPV) of 96.87%, and an area under the receiver operating

characteristic curve (AUROC) of 0.91. These results underscore the potential of deep learning to deliver superior predictive performance compared to traditional ML models, particularly in complex clinical settings.

Despite the growing evidence supporting DL models, their adoption in medicine remains limited due to concerns about their interpretability. As with other AI applications, DL models are often perceived as "black boxes" because they do not inherently provide insights into how predictions are made (Cabitza et al., 2017). This is particularly problematic in high-stakes fields such as surgery, where understanding the underlying risk factors that contribute to a prediction is essential for clinicians. Techniques like Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are crucial tools that can help make machine learning models more interpretable by identifying key features driving their predictions (Lundberg & Lee, 2017; Ribeiro et al., 2016). By making these models more transparent, clinicians can gain greater trust in their predictive outputs and better incorporate them into surgical decision-making processes.

Lee et al. (2018) developed a deep neural network model trained on 87 intraoperative features to predict postoperative in-hospital mortality. These features were calculated or extracted automatically at the end of the surgery, aiming to offer real-time predictive insights. The study compared the deep neural network's performance to published clinical risk scores, administrative risk scores, and a logistic regression model using the same intraoperative features. While the deep neural network demonstrated the ability to predict in-hospital mortality, the authors concluded that it was not yet superior to existing clinical methods. This finding highlights both the potential and the current limitations of deep learning models in the medical field, emphasizing the need for



continued refinement and validation in diverse clinical settings. Deep learning (DL) models have gained significant attention due to their ability to operate without human-designed rules, relying instead on large datasets to map inputs to specific outputs. One of the most advantageous features of DL is its capacity for automatic feature extraction, which contrasts with traditional methods that often require extensive domain expertise to handcraft features. As Alzubaidi et al. (2021) highlight, DL algorithms achieve feature extraction automatically, minimizing human effort and the need for field knowledge. This has motivated researchers to leverage DL in tasks where extracting discriminative features is complex, such as predicting postoperative mortality, where nuanced patterns in data may be better captured by machine learning rather than human intuition. Health care has the problem of collection of large amounts of data due to privacy issues. This raises two questions: first, how can we compensate for small datasets to make them suitable for deep learning? Second, how can we address the issue of highly imbalanced data, particularly for targets such as mortality and complications? .Unequal distribution of classes in the training data set leads to poor predictive performance, particularly for the minority class. Classifiers trained on imbalanced data tend to be biased toward the majority class, resulting in higher classification errors for the minority class (Tomescu, Czibula, and Nițică, 2021). This is a common issue in clinical prediction models, where the class of interest (e.g., patients at high risk of mortality) often forms a small proportion of the overall dataset. To address this, deep generative models have shown remarkable potential in generating highly realistic content, such as images, texts, and sounds, and they can be similarly leveraged to create synthetic data to balance class distributions (Goodfellow et al., 2014). This capability offers a promising solution for overcoming the challenges posed by small, imbalanced datasets. Additionally, one of the major barriers to the broader adoption of machine learning in healthcare is the “black box” nature of many models. Both clinicians and patients often

struggle to understand how predictions are made, which raises concerns about trust and transparency (Lipton, 2017). This underscores the need for explainable and interpretable models in clinical settings to ensure that decisions based on machine learning can be understood and validated by medical professionals, facilitating greater acceptance and utilization of these technologies in healthcare.

### **1.3 Purpose of research**

#### **1.3.1. Overall Objective**

1. To create a deep learning model for forecasting mortality and assess its performance against various machine learning methods, with the goal of adopting the most effective model to aid surgeons in their decision-making process.
2. To identify significant variables using model-agnostic interpretability methods such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP).

#### **1.3.2 Specific Aim**

To explore whether Variational Autoencoder (VAE) can augment the small data set of patients who underwent general surgical procedures in a deep learning model and correct for imbalance in training data with superior results.?

### **1.4 The significance of the study**

The significance of this study lies in the development of a calibrated probabilistic model that will provide accurate predictions of surgical outcomes. Once deployed, this model will assist surgeons in counselling patients and their families about the likely results of surgery, helping to inform decision-making, manage expectations, and improve overall patient care by offering personalized risk assessments. The model's ability to predict outcomes reliably will be especially valuable in enhancing preoperative discussions and supporting clinical decision-making.

## **1.5 RESEARCH PURPOSE AND QUESTION/HYPOTHESIS**

The purpose of this research is to develop a robust and interpretable deep-learning model for forecasting mortality among surgical patients, providing surgeons with a reliable tool to support critical decision-making. The study aims to create a model that not only achieves high predictive accuracy but also demonstrates its superiority over traditional machine learning techniques, identifying the most effective model for real-world application in a surgical setting. Additionally, by employing model-agnostic interpretability methods such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), the research seeks to uncover key variables influencing patient outcomes. This approach not only enhances model transparency but also highlights clinically relevant factors, facilitating a deeper understanding of the determinants of surgical mortality and improving trust in the model's predictions among healthcare professionals. Through these objectives, the study aspires to bridge advanced machine learning methodologies with clinical applicability, contributing a valuable tool for informed decision-making in healthcare.

**RQ1** Which approach, Machine Learning or Deep Learning, is more effective for predicting mortality in a small dataset supplemented with synthetic data

**RQ2** Whether Generational autoencoder can be used using variational autoencoder to correct for imbalance in training data with superior results?

**RQ3.** Can the variables of importance identified by the explainability techniques, Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP), provide consistent results in explaining and interpreting the model for predicting postoperative mortality?

## **CHAPTER II: REVIEW OF LITERATURE**

### **2.1 Introduction & Theoretical Framework**

The Charlson and Elixhauser comorbidity classification systems are the most frequently utilized in health research. According to Van Walraven et al. (2009), these systems, established by Charlson in 1994 and Elixhauser in 1998, are significantly linked to various outcomes, including in-hospital mortality (Poses et al., 1996; Sundararajan et al., 2004) and all-cause mortality after discharge, as noted by Charlson et al. (1994). When directly compared, research indicates that the Elixhauser comorbidity system is marginally more effective than the Charlson system in accounting for comorbidity (Southern, Quan, and Ghali, 2004; Stukenborg, Wagner, and Connors, 2001; Dominick et al., 2005; Lieffers et al., 2011). Despite this, the Charlson comorbidity index remains in use (Yu et al., 2003; Kil et al., 2012; Dailiana et al., 2013; Bannay et al., 2016; Lunde et al., 2019; Briongos-Figuero et al., 2020). In adult population-based cohorts, Simard et al. (2018) demonstrated that combining the Elixhauser and Charlson indices yields better results for predicting 30-day mortality than using either index alone. Gagne et al. (2011) also found that the Combined Comorbidity Score, which integrates conditions from both Charlson and Elixhauser, was more accurate in predicting mortality among Medicare patients. However, Van Walraven et al. (2009) noted that the Elixhauser comorbidity score's poor performance in predicting hospital mortality rates is not surprising, as comorbidity is just one of several factors that significantly affect the risk of in-hospital death. Escobar et al. (2008) identified

four other factors—urgency of admission, service type, patient age, laboratory abnormalities, and admission diagnosis—that were as influential, if not more so, in predicting hospital death risk as comorbidities. Pine et al. (2007) highlighted the importance of incorporating 'present on admission' (POA) comorbidity codes and laboratory test results from the first 24 hours in the hospital when comparing inpatient mortality for certain conditions. In another study, Smith et al. (1991) examined the Health Care Financing Administration's (HCFA) use of hospital billing data—such as age, sex, and diagnoses—to develop statistical models for estimating the risk of death during and after hospital stays. They compared these models to severity classifications and clinical risk factors for mortality prediction. Their findings indicated that the inclusion of clinical risk factors leads to more accurate death risk estimations and provides a better measure of care quality. Hanan et al. (1992) analyzed data from New York State's Cardiac Surgery Reporting System, which records cardiac preoperative risk factors, postoperative complications, and hospital discharge information. Their study aimed to identify significant clinical risk factors and pinpoint cardiac surgery centers with potential quality-of-care issues. However, Pine et al. (1998) found that inpatient mortality prediction significantly improved when laboratory values were combined with administrative data that included only secondary diagnoses present on admission (i.e., comorbidities). They observed that adding further clinical data contributed little to predictive accuracy. Pine et al. also highlighted that risk adjustment methods that included conditions developed during hospitalization provided better mortality predictions than models focused only on admission diagnoses. However, they argued that including all diagnoses might undermine

the goal of adjusting for the patient's condition at the start of care. Risk adjustment models that claim to measure a patient's illness severity at admission but also include fatal hospital-acquired complications, such as cardiac arrest, shock, or hypotension, may obscure poor care by artificially inflating the perceived riskiness of patients who worsen during hospitalization. Glance et al. (2006) demonstrated that including complications in risk adjustment models can unfairly benefit hospitals with poor performance, as it may give them "credit" for their complications, potentially misclassifying them as delivering better care than they truly do. However, as noted by the previously mentioned authors, if the goal of the model is to predict patient mortality, it is reasonable to incorporate comorbidities, laboratory results, and any complications that arise during hospitalization. Numerous perioperative risk scores and prediction models have been developed globally (Moonesinghe et al., 2013), such as the American Society of Anaesthesiologists-Physical Status (ASA-PS) (Saklad, 1941), the Physiological and Operative Severity Score for the enumeration of Mortality and Morbidity (POSSUM) (Copeland, Jones, and Walters, 1991), the Surgical Outcome Risk Tool (SORT) (Protopapa et al., 2014), and the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) (Bilimoria et al., 2013; Chiew et al., 2020). However, each of these tools has its limitations. For instance, the ASA-PS suffers from high inter-user variability (Cohen et al., 2009), POSSUM requires data not typically available preoperatively (Copeland, Jones, and Walters, 1991; Protopapa et al., 2014; Chiew et al., 2020; Cohen et al., 2009; Brooks, Sutton, and Sarin, 2005), and both SORT (Protopapa et al., 2014) and ACS-NSQIP (Chiew et al., 2020) lack external validation outside their original populations. Additionally, ACS-

NSQIP is criticized for the complexity of its model (Reilly, 2021). These risk-scoring systems are primarily based on statistical techniques like logistic or Cox regression, which remain the standard for developing prediction models, even when the primary focus is accuracy over the interpretation of the regression coefficients (Knaus et al., 1991; Le et al., 1993). Most of these tools are designed for preoperative risk assessment and thus rely only on variables available before surgery (Bilimoria et al., 2013). However, the patient's operative procedure and complications in the immediate aftermath of surgery provide additional information for predicting mortality and potentially permitting early recognition of modifiable risk factors. Machine learning, often referred to as data mining, focuses on discovering meaningful patterns in data to address specific questions and may provide enhanced predictive performance, while also enabling automation and integration into clinical decision support systems (Taylor et al., 2016). It shares close ties with both statistics and engineering (Wu et al., 2010). Although machine learning algorithms have the potential to improve prediction accuracy compared to traditional regression models by capturing complex, nonlinear relationships within data, they cannot extract information that does not already exist in the dataset, regardless of the sophistication of the algorithm or computing power (Chen and Asch, 2017).

In recent years, machine learning (ML) has gained widespread adoption in mortality and disease prediction, demonstrating superior classification and predictive performance with large datasets (Taylor et al., 2016; Rose, 2013; Lien et al., 2021; Rau et al., 2019; Delahanty et al., 2018; Fleuren et al., 2020). Traditional ML methods often rely on hand-engineered features, where each predictive outcome requires the creation of a

custom dataset with specific variables. The effectiveness of ML models is largely influenced by the quality of the data representation (or features) used (Bengio, Courville, and Vincent, 2013). As a result, a significant amount of effort in deploying ML algorithms is focused on designing preprocessing pipelines and data transformations to create representations that enable effective learning. This feature engineering process, while crucial, is time-consuming and emphasizes a key limitation of current algorithms: their inability to autonomously extract and structure relevant information from the data. In contrast, deep learning methods are recognized for their ability to automatically derive meaningful features from raw data (Cosgriff and Celi, 2020). A brief overview of various approaches is provided below:

### **2.1.1 Machine Learning Techniques**

In constructing a machine learning system, three core components are essential: data, models, and the learning process (Deisenroth, Faisal, & Ong, 2020). At the heart of this system is the quest to determine what constitutes an effective model, and while the definition may vary with data and applications, a universally accepted principle is that a strong model should generalize well to new, unseen data. Establishing meaningful performance metrics—like accuracy or error rates—helps benchmark model effectiveness, ensuring that predictions approximate real outcomes. By focusing on automation, machine learning leverages algorithms to draw critical insights from data, thus providing adaptable solutions across a range of datasets without requiring extensive domain-specific knowledge (Deisenroth, Faisal & Ong, 2020).



Machine learning aims to devise universal methods for identifying key patterns within data, which is why data itself forms the foundation of the process. In practice, a model serves as an approximation of the underlying mechanism generating the observed data, designed to capture essential characteristics and uncover latent patterns. A well-built model enables the prediction of real-world events, potentially circumventing the need for actual experimentation (Deisenroth, Faisal & Ong, 2020). With a reliable model structure, one can draw valuable insights, furthering machine learning's overarching goal of providing predictive power without exhaustive data collection.

The learning phase in machine learning is critical. Given a dataset and a model framework, the objective is for the model to succeed on previously unseen data. This requirement distinguishes learning from mere memorization; a model focused solely on training data risks overfitting, leading to poor generalization. Hence, effective learning focuses on achieving configurations that perform optimally across a broader data landscape. In practical workflows, machine learning can be broken down into three key stages: prediction (or inference), training (or parameter estimation), and hyperparameter tuning (or model selection). The prediction stage applies the model to new data, with probabilistic models particularly benefiting from uncertainty quantification, a feature that aids in assessing predictive confidence. Incorporating probability theory, this step allows models to express the confidence level of each prediction based on the data context (Deisenroth, Faisal & Ong, 2020).

Training focuses on optimizing parameters for maximal performance, with most techniques rooted in gradient-based methods that refine model accuracy. This optimization

is akin to ascending a hill, where reaching the peak equates to obtaining the best parameter configuration (Deisenroth, Faisal & Ong, 2020). For non-probabilistic models, the Empirical Risk Minimization (ERM) principle addresses overfitting by minimizing empirical risk; this approach optimizes the function to fit the training dataset closely, but overfitting often occurs when the empirical risk underestimates the true risk on new data (Deisenroth, Faisal & Ong, 2020).

A machine learning system is fundamentally constructed around three essential components: data, models, and the process of learning (Deisenroth, Faisal, & Ong, 2020). A critical question in the field is, "What makes a model effective?" While defining a "good" model can be complex due to the nuances and variability of data, a core principle is that a robust model should generalize effectively to new, unseen data. To meet this criterion, it is necessary to establish clear performance metrics, such as accuracy or error rates compared to actual outcomes, and to fine-tune models to excel according to these standards. Machine learning ultimately focuses on developing algorithms that automatically derive significant insights from data, making "automation" a central theme (Deisenroth, Faisal & Ong, 2020). In this sense, machine learning is designed to be versatile, producing methods applicable across various datasets and yielding meaningful results without extensive domain-specific expertise.

The primary aim of machine learning is to develop general-purpose methods that identify important patterns within data, with data being fundamental to the process. A model in machine learning typically seeks to replicate the underlying process that generates data similar to the dataset in question. An effective model is a simplified representation of the

data-generating process, capturing essential features needed to reveal patterns within the data. With a well-constructed model, predictions about real-world scenarios become feasible, eliminating the need for real-world experimentation to achieve similar outcomes (Deisenroth, Faisal & Ong, 2020).

Learning, the third key component, is crucial. Given a dataset and a model framework, the primary goal is for the model to perform well on new data it has not encountered before, rather than just memorizing patterns from the training data. A model that performs well solely on training data may be overfitting, which often fails to translate to strong results on unseen data. In practical applications, machine learning models must adapt to scenarios they haven't previously encountered. Thus, the learning goal is to identify an optimal model configuration and its parameters to maximize performance on unseen data. In practice, three phases typically define machine learning workflows:

1. **Prediction or inference**
2. **Training or parameter estimation**
3. **Hyperparameter tuning or model selection**

During the prediction phase, a model is applied to new data to make inferences based on learned parameters. This phase differs based on whether the model is deterministic or probabilistic. When probabilistic models are involved, the prediction phase is referred to as inference, as it incorporates uncertainty into predictions.

Uncertainty quantification is another important aspect of machine learning, allowing practitioners to gauge the confidence of predictions at specific data points. Probability theory underpins this uncertainty assessment, providing a theoretical framework for

expressing predictive confidence. In model training, the objective is generally to adjust parameters to maximize performance. Many training techniques depend on gradient-based optimization, which directs the adjustments needed to improve model accuracy. Training a model, therefore, involves refining its parameters based on a utility function that measures the model's fit to the data. This optimization can be likened to climbing a hill, where the summit represents the optimal parameter set for the model (Deisenroth, Faisal & Ong, 2020).

In the training or parameter estimation phase, we adjust our predictive model based on training data to identify effective predictors. Two primary strategies are employed in this process: finding the best predictor based on a measure of quality, often referred to as a *point estimate*, or utilizing *Bayesian inference* (Deisenroth, Faisal & Ong, 2020). While both strategies can apply to different types of predictive models, Bayesian inference specifically requires probabilistic models (Deisenroth, Faisal & Ong, 2020).

### **Non-Probabilistic Models**

For *non-probabilistic models*, we typically follow the principle of **Empirical Risk Minimization (ERM)**. This principle frames an optimization problem aimed at minimizing the empirical risk

$$R_{emp}(f; D) = \frac{1}{n} * \sum_1^n \ell(f(x_i), y_i)$$

- $D = (f(x_i), y_i)_{i \text{ to } n}$   
*is the training dataset.*
- $f$  *is the predictive function.*

- $\ell$  is the loss function that measures the difference between the predicted values  $f(x_i)$  and the true values  $y_i$ .

(Deisenroth, Faisal & Ong, 2020)

This minimization allows us to find parameters that provide good predictions. However, it can lead to overfitting, where the model learns the training data too closely and fails to generalize well to new data. This phenomenon often occurs when the empirical risk on the training set  $R_{\text{emp}}(f; D_{\text{train}})$  significantly underestimates the true risk  $R_{\text{true}}(f)$  (Deisenroth, Faisal & Ong, 2020).

### Maximum Likelihood Estimation

In the context of statistical models, the principle of Maximum Likelihood Estimation (MLE) is utilized to find a good set of parameters  $\theta$ . MLE seeks to maximize the likelihood function

$$L(\theta; D) = \prod_{i=1}^n p(y_i | x_i, \theta)$$

where:

- $p(y_i | x_i, \theta)$  is the probability of observing  $y_i$  given  $x_i$  and parameters  $\theta$  (Deisenroth, Faisal & Ong, 2020).

Maximizing the likelihood is equivalent to minimizing the negative log-likelihood  $L(\theta)$

$$L(\theta) = - \sum_{i=1}^n \log p(y_i | x_i, \theta)$$

This approach provides a powerful means of fitting statistical models to data, although it does not inherently account for uncertainty in the model's predictions or parameters (Deisenroth, Faisal & Ong, 2020).

## Probabilistic Models

Focusing solely on some statistic of the posterior distribution (such as the parameters  $\theta$  that maximizes the posterior) leads to a loss of information which can be critical in a system that uses the prediction  $p(x_i/\theta)$  to make decisions. These decision-making systems typically have different objective functions than the likelihood of squared-error loss or a misclassification error. Therefore, having the full posterior distribution around can be extremely useful and leads to more robust decisions. *Bayesian inference* is about finding this posterior distribution (Gelman et al., 2004). For a dataset  $X$ , a parameter prior  $p(\theta)$ , and a likelihood function, the posterior

$$p(\theta | X) = \frac{p(X | \theta)p(\theta)}{p(X)}$$

is obtained by applying Bayes' theorem. The key idea is to exploit the Bayes theorem to invert the relationship between the parameters  $\theta$  and the data  $X$  (given by the likelihood) to obtain the posterior distribution  $p(\theta | X)$ . The implication of having a posterior distribution on the parameters is that it can be used to propagate uncertainty from the parameters to the data. More specifically, with a distribution  $p(\theta)$  on the parameters our predictions will be For *probabilistic models*, **Bayesian inference** is used to model uncertainty. This approach incorporates:

- **Prior**  $p(\theta)$ : Represents initial beliefs about the parameter values before observing any data.
- **Likelihood**  $p(y | x, \theta)$ : The probability of observing data  $y$  given input  $x$  and parameters  $\theta$

- **Posterior  $p(\theta | y)$** : The updated beliefs about the parameters after observing data  $y$  (Deisenroth, Faisal & Ong, 2020).

According to Bayes' theorem, the posterior can be expressed as:

$$p(\theta | y) = p(y | x, \theta) * p(\theta) / p(y)$$

For optimization purposes, since  $p(y)$  does not depend on  $\theta$ , we can simplify this to:

$$p(\theta|y) \propto p(y|x,\theta) \cdot p(\theta)$$

This formulation illustrates how prior knowledge can be integrated with observed data to refine parameter estimates. The process of estimating the parameters by maximizing the posterior distribution is known as **Maximum A Posteriori (MAP) estimation** (Deisenroth, Faisal & Ong, 2020).

### **Regularization**

In non-probabilistic models, regularization techniques are often introduced to prevent overfitting. Regularization adds a penalty term  $\Omega(\theta)$  to the loss function:

$$R_{reg}(f; D) = R_{emp}(f; D) + \lambda \Omega(\theta)$$

where:

- $\lambda$  controls the strength of the penalty,
- $\Omega(\theta)$  is a regularization function (e.g., L1 or L2 regularization) (Deisenroth, Faisal & Ong, 2020).

In probabilistic models, the concept of regularization is analogous to the **prior distribution** on the parameters  $p(\theta)$ , which biases the parameter estimates toward simpler models (Deisenroth, Faisal & Ong, 2020).

Thus, in short, **non-probabilistic models** primarily use **MLE** and **empirical risk minimization**, focusing on optimization problems to reduce loss without explicitly modeling parameter uncertainty (Deisenroth, Faisal & Ong, 2020).

**Probabilistic models** employ **Bayesian inference**, incorporating priors, likelihoods, and posteriors to comprehensively represent uncertainty (Deisenroth, Faisal & Ong, 2020).

Each approach has its advantages, with non-probabilistic methods being straightforward for optimization tasks, while probabilistic models excel in situations that require uncertainty quantification (Deisenroth, Faisal & Ong, 2020). Thus, Model parameters can be estimated using either maximum likelihood estimation (MLE) or maximum a posteriori (MAP) estimation. Both methods yield a single best estimate of the parameter, making parameter estimation primarily an optimization problem. Once these parameters are estimated, they can be utilized for making predictions. Specifically, the predictive distribution takes the form  $p(x | \theta)$  where the estimated parameters are applied within the likelihood function. However, relying only on a point estimate from the posterior distribution (such as the parameter  $\theta$  that maximizes the posterior) may result in a loss of valuable information. This can be especially problematic in decision-making systems where predictions, like  $p(x_i | \theta)$ , are used. These systems typically have objective functions beyond just likelihood, such as minimizing squared-error loss or



misclassification errors. Hence, retaining the full posterior distribution can yield more robust decisions by providing a richer uncertainty measure (Gelman et al., 2004).

In Bayesian inference, the posterior distribution is derived for a given dataset  $X$ , prior  $p(\theta)$ , and likelihood function, using Bayes' theorem:

$$p(\theta | X) = p(X | \theta)p(\theta)/p(X), p(X) = \int p(X | \theta)p(\theta)d\theta.$$

The essence of Bayesian inference is to apply Bayes' theorem to reverse the relationship between the parameters  $\theta$  and the data  $X$  (as described by the likelihood function), resulting in the posterior distribution  $p(\theta | X)$ . The benefit of obtaining a posterior distribution for the parameters is that it allows for uncertainty in parameter estimates to be reflected in the predictions. By having a parameter distribution  $p(\theta)$ , the predictions are expressed as

$$p(x) = \int p(x | \theta)p(\theta)d\theta = E_{\theta}[p(x | \theta)],$$

where predictions are averaged over all plausible parameter values  $\theta$ , with each value's plausibility determined by  $p(\theta)$

Comparing these approaches, parameter estimation through MLE or MAP provides a point estimate  $\theta$  for the parameters, requiring optimization as the central computational step. On the other hand, Bayesian inference produces a distribution (posterior), making integration the core computational challenge. Predictions based on point estimates are direct, while predictions in a Bayesian context involve another integration to account for parameter uncertainty; Bayesian inference, however, offers a structured approach for incorporating prior knowledge, side information, and model structure, which is often challenging in conventional parameter estimation. Furthermore, transferring parameter uncertainty into

predictions can enhance decision-making, which is particularly beneficial in risk assessment and exploration for data-efficient learning (Deisenroth et al., 2015; Kamthe & Deisenroth, 2018).

### **Deterministic Loss Function**

The mean squared error (MSE) is one of the simplest and most widely used loss functions, especially for regression tasks, as it provides a clear measure of prediction error. Here, MSE is used to calculate loss and adjust the weights of deterministic models accordingly:

$$MSE = 1/N \sum_{i=1}^n (y^i - y_i)^2$$

where N is the number of predictions,  $y^i$  represents the predicted data, and  $y_i$  represents the true label data. MSE calculates the difference between each prediction and the actual value, squares it to avoid negative values, and then averages these squared errors. This result provides a single MSE value, indicating how far or close the model's predictions are from the real values (Goodfellow et al., 2016).

When dealing with binary outcomes, a common choice for the loss function is **binary cross-entropy** (or log loss), as it effectively measures the performance of a model where each prediction is a probability between 0 and 1. Binary cross-entropy is especially suitable for classification tasks with probabilistic outputs, where it penalizes incorrect predictions with an exponential increase in error for increasingly incorrect probabilities. The binary cross-entropy loss L is defined as follows:

$$-1/N \sum_{i=1}^n [y_i \log(y^i) + (1 - y_i) \log(1 - y^i)]$$

where:

- N is the number of predictions,
- $y_i$  is the true binary label (0 or 1),
- $y^i$  is the predicted probability of the positive class (i.e., the probability that  $y_i=1$ ).

In this function, when the prediction aligns perfectly with the true label (e.g.,  $y_i=1, y^i=1$  and  $y^i$  is close to 1), the loss value is minimized. Conversely, if the prediction diverges from the true label, the loss increases. This nature of binary cross-entropy makes it highly effective for tasks like binary classification in deep learning and is commonly used in probabilistic models as well as traditional logistic regression (Goodfellow et al., 2016).

### **Practical Applications**

Binary cross-entropy is critical in applications where the outcome is binary, such as medical diagnosis (e.g., disease present vs. absent), financial forecasting (e.g., market up or down), and other binary decision-making tasks (Chollet, 2018). In Bayesian neural networks, this function is often combined with probabilistic estimations, allowing models to produce not only a binary decision but also a measure of uncertainty.

### **Bayesian Loss Function**

For Bayesian models, the loss function is more complex. Starting with the posterior distribution  $P(w|D)$  over the model's parameters, the direct calculation often lacks a closed-form solution (Wen et al., 2018). To approximate this, variational inference is used, transforming the problem into one of finding a distribution  $q_\theta(w)$  that is close to  $P(w|D)$ , with similarity measured by the Kullback-Leibler (KL) divergence:

$$KL(q\theta(w) \parallel P(w \mid D)) = \int q\theta(w) \log P(w \mid D) q\theta(w) dw$$

This setup allows us to approximate the posterior distribution by minimizing the KL divergence between our model's parameter distribution and the target posterior. However,  $P(w|D)$  is still difficult to calculate directly, so we use the Evidence Lower Bound (ELBO) to make this optimization computationally feasible (Chollet, 2018). The ELBO reformulates the objective, allowing the optimization problem to be reframed as maximizing ELBO:

$$\theta^* = \arg \theta \max \int q\theta(w) \log q\theta(w) P(D \mid w) P(w) dw$$

Maximizing this expression provides an efficient way to optimize the Bayesian loss function and train Bayesian neural networks.

### 2.1.1.1 Tree-based models: Decision Tree

Decision trees divide sample data by splitting variables at specific points and are often visualized as a tree structure (Kuhn & Johnson, 2013; Hastie et al., 2009). Some well-known decision tree algorithms include Quinlan's ID3, C4.5, C5 (Quinlan, 1979, 1983, 1993), and CART (Breiman et al., 1984). CART (Classification and Regression Trees) handles both classification and regression tasks. In CART, tests are always binary, and it uses the Gini diversity index to rank tests. Trees are pruned using a cost-complexity model, with parameters estimated through cross-validation. The Gini rule, favored by the CART authors, is similar to the more widely known entropy or information-gain criterion.

For a binary (0/1) target, the "Gini measure of impurity" at node  $t$  is defined as:

$$G(t) = 1 - p(t)^2 - (1 - p(t))^2$$

where  $p(t)$  is the (possibly weighted) relative frequency of class 1 at the node. The improvement (gain) from splitting a parent node  $P$  into left and right children  $L$  and  $R$  is:

$$I(P) = G(P) - qG(L) - (1-q)G(R)$$

where  $q$  is the (possibly weighted) fraction of instances going to the left. CART favors the Gini criterion because it is more computationally efficient than information gain and can be extended to include symmetric costs. Later versions of CART added information gain as an optional splitting rule. CART also introduced the modified towing rule, which compares the target attributes directly. For regression (continuous targets), CART offers Least Squares (LS) and Least Absolute Deviation (LAD) criteria to measure split improvements (Wu et al., 2010). Three machine learning algorithms—Random Forest (RF), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGB)—use decision trees as their base learner (Schapire, 2003).

#### **2.1.1.2 Random Forest (RF)**

Because decision trees often have limited predictive accuracy, several methods are used to combine multiple trees to enhance performance. These methods include bagged trees and boosted trees (Churpek et al., 2016). Random Forest (RF), a bagged tree model (Breiman, 2001), is a classification and regression method that aggregates the predictions from a large number of decision trees. Specifically, RF builds an ensemble of trees from subsamples of the training dataset, internally validating them to predict the response based on the predictors. Each tree is a standard Classification or Regression Tree (CART) that uses the Decrease of Gini Impurity (DGI) as the splitting criterion, selecting a predictor from a randomly chosen subset of variables (different at each split). Each tree is constructed from

a bootstrap sample drawn with replacement from the original dataset, and predictions from all trees are combined using majority voting.

One key feature of RF is the out-of-bag (OOB) error. For each tree, some observations are not used in the training process, making them OOB observations, which serve as an internal validation set. The OOB error is the average error frequency when OOB observations are predicted using the trees that did not include them during training, providing a less optimistic and reliable estimate of the error for independent data.

RF requires setting two key parameters: the number of trees (ntree) and the number of randomly selected predictor variables (mtry).

Pera et al. (2022) developed a 90-day mortality (90DM) risk prediction model using machine learning in a large multicentre cohort of patients undergoing gastric cancer resection with curative intent. They tested four 90DM predictive models based on preoperative clinical characteristics: Cross-Validated Elastic Net regularized logistic regression (cv-Enet), boosting linear regression (glmboost), random forest, and an ensemble model. Among the single models, RF showed the best discrimination ability, with a validated AUC of 0.844 (95% CI: 0.841–0.848), outperforming cv-Enet (AUC of 0.796, 95% CI: 0.784–0.808) and glmboost (AUC of 0.797, 95% CI: 0.785–0.809). The ensemble model did not significantly improve the AUC (0.847, 95% CI: 0.836–0.858) compared to the RF model alone.

The study, however, did not give details of other metrics including precision, recall and F1 score. For ‘‘imbalanced’’ datasets, a metric such as the area under the precision-recall curve - plotting the positive predictive value (precision) against sensitivity (recall) - is often

more informative, particularly to quantify the presence of false alarms. For the clinician at the bedside, a model with a high false alarm rate is unlikely to be a useful model. In addition, if a false positive decision causes greater harm than a false-negative decision, a model with high specificity may be preferable to a model with high sensitivity and lower specificity, although the latter model might have, say, a higher AUROC (Vistisen et al.,2022). While random forests result in more reliable predictions than single trees, they are difficult to interpret as individual trees are lost in the overall forest. To achieve visual interpretability, a single tree most similar to the overall forest can be identified and extracted (Chirikov et al.,2017). Churpek et al.(2016) in a Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration including mortality in the wards found that the random forest model was the most accurate (AUC, 0.80 [95% CI, 0.80-0.80]) followed by the gradient boosted machine (AUC, 0.80 [95% CI, 0.79–0.80]). The logistic regression model with spline terms was more accurate than the model utilizing linear predictor terms (AUC, 0.77 vs 0.74;  $p < 0.01$ ), and all models were more accurate than the Modified Early Warning Score (AUC, 0.70 [95% CI, 0.70-0.70]), a commonly utilized rapid response team activation tool. One-year mortality was 11% ( $n = 3738$ ). The comparison with the Modified Early score displayed that nonlinear models are more accurate. However, the authors did not use Deep learning despite having a high volume of data which may have given better insight about nonlinear ML techniques. Hill et al.(2019) reported on the use of machine learning algorithms, specifically random forests, to create a fully automated score that predicts postoperative in-hospital mortality based solely on structured data available at the time of surgery. They used electronic health

record data from 53 097 surgical patients (2.01% mortality rate) who underwent general anesthesia between April 1, 2013, and December 10, 2018, in a large US academic medical centre to extract 58 preoperative features. The model was created using a set of features including basic patient information such as age, sex, BMI, BP, and HR; laboratory tests frequently obtained before surgery, such as sodium, potassium, creatinine, and blood cell counts; and surgery-specific information such as the surgical procedure codes. In total, 58 preoperative features (including ASA status) were selected. For all variables, only the most recent value before surgery was included. Authors evaluated four different classification models: logistic regression, Elastic Net logistic regression, random forests, and gradient-boosted trees. A random forest classifier with area under the curve [AUC] of 0.932, 95% confidence interval [CI] 0.910e0.951) outperformed Preoperative Score to Predict Postoperative Mortality (POSPOM) scores (AUC of 0.660, 95% CI 0.598e0.722), Charlson comorbidity scores (AUC of 0.742, 95% CI 0.658e0.812), and ASA physical status (AUC of 0.866, 95% CI 0.829e0.897). Including the ASA physical status with the preoperative features achieved an AUC of 0.936 (95% CI 0.917e0.955). Gradient boosting trees can be more accurate than random forests because the model includes training the trees to correct each other's errors. Hence, it can capture complex patterns in the data. However, if the data are noisy, the boosted trees may overfit and start modeling the noise. RF has only one hyperparameter to set: the number of features to randomly select at each node. However, there is a rule of thumb to use the square root of the number of total features which works pretty well in most cases (Bernard, Heutte, and Adam,2009). On the other hand, GBMs have several hyperparameters that include the number of trees,



the depth (or number of leaves), and the shrinkage (or learning rate). While it is not true that RF does not overfit (as opposed to what many are led to believe by Breiman's strong assertions), it is true that they are more robust to overfitting and require less tuning to avoid it.

### **2.1.1.3 Boosting algorithm**

The principal difference between boosting and the committee methods such as bagging discussed above, is that the base classifiers are trained in sequence, and each base classifier is trained using a weighted form of the data set in which the weighting coefficient associated with each data point depends on the performance of the previous classifiers. In particular, points that are misclassified by one of the base classifiers are given greater weight when used to train the next classifier in the sequence. Once all the classifiers have been trained, their predictions are then combined through a weighted majority voting scheme (Bishop and Nasrabadi,2006). Originally designed for solving classification problems, boosting can also be extended to regression (Friedman,2001). The most widely used form of boosting algorithm called AdaBoost, short for 'adaptive boosting', was developed by Freund and Schapire (1996). Extreme Gradient Boosting (XGBoost or XGB for short) is an optimized implementation of a GBM (Chen and Guestrin,2016). It uses decision (regression) trees as weak learners. To perform the gradient descent procedure, it calculates the loss and adds a tree to the model (always one at a time) that reduces it (i.e., follows the gradient). This is done by parameterizing the tree and modifying these parameters to move in the right direction by reducing the loss. The existing trees in the model are not changed. Trees are added until a fixed number is reached, until the loss

reaches an acceptable level, or until no more improvement is achieved. XGB's final output is given by the (weighted) sum of all the predictions made by all the individual trees. Barash et al. (2022) developed an ML model for predicting 30-day mortality in patients discharged from the Emergency Department (ED). The overall rate of 7-day post-ED discharge mortality was 571/363 635 (0.2%), and the 30-day mortality rate was 2989/363 635 (0.8%). A gradient-boosting model was trained to predict mortality within 30 days of release from the ED. The default XGBoost parameters: eta=0.3, max depth=6, and scale pos weight=1 was used. Estimators were set to 1000. The class imbalance was addressed by the XGBoost class weights feature. Missing values were modeled by the XGBoost algorithm. Bootstrapping validations (1000 bootstrap resamples) were used to calculate 95% CI. As reported by the authors, for the entire cohort, the gradient boosting model showed an AUC of 0.97 (95% CI 0.96 to 0.97). For a fixed specificity of 95% and a false positive rate (FPR) of 1:20, the model showed a sensitivity of 84% for oncology patients. Other matrices had a Positive predictive value of 0.18 (95% CI 0.17 to 0.19), Negative predictive value of 1.00 (95% CI 1.00 to 1.00), F1 score of 0.29 (95% CI 0.27 to 0.30). For the non-oncological cohort model for a fixed specificity of 95% and FPR 1:20, the sensitivity of this model was 74%. As can be seen, while the model has a high AUC, the F1 score is less due to low PPV. Chiew et al. (2020) compared the performance of machine learning models against the traditionally derived Combined Assessment of Risk Encountered in Surgery (CARES) model and the American Society of Anesthesiologists Physical Status (ASA-PS) in the prediction of 30-day postsurgical mortality and need for intensive care unit (ICU) stay >24 hours. Candidate models were trained using random forest, adaptive boosting, gradient

boosting, and support vector machine. Models were evaluated on the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). Gradient boosting was the best-performing model with an F1 score of 0.28 and AUPRC of 0.23 and 0.38 for mortality and ICU admission outcomes respectively. Forte et al. (2017) developed ML algorithms to predict 5 yrs., long-term mortality in Coronary Artery Bypass Graft (CABG) patients using data from routinely measured clinical parameters from a large cohort of CABG patients (n=5868) and compared the accuracy of 5 different ML models with traditional Cox and Logistic Regression. In the validation dataset, the Gradient Boosted Machine (GBM) algorithm was the most accurate (AUROC curve [95%CI] of 0.767 [0.739-0.796]), proving to be superior to traditional Cox and logistic regression ( $p < 0.01$ ) for long-term mortality prediction. However, the only metric used by authors was ROC value which does not provide enough information in imbalanced data for adequate comparison in terms of False positives and False negatives. The authors did not use deep learning for comparison. Peng et al. (2022) developed machine learning models to predict postoperative major adverse cardiovascular events in geriatric patients. They trained various models, including extreme gradient boosting (XGB), gradient boosting machine, random forest, support vector machine, and Elastic Net logistic regression. The performance of these models was compared using the area under the precision-recall curve (AUPRC), the area under the receiver operating characteristic curve (AUROC), and the Brier score. XGB outperformed the other models, achieving an AUPRC of 0.404 (95% CI: 0.219–0.589), an AUROC of 0.870 (95% CI: 0.786–0.938), and a Brier score of 0.024 (95% CI: 0.016–0.032). The model trained on an under-sampled

dataset showed even better performance, with an AUPRC of 0.511 (95% CI: 0.344–0.667,  $p < .001$ ), an AUROC of 0.912 (95% CI: 0.847–0.962,  $p < .001$ ), and a Brier score of 0.020 (95% CI: 0.013–0.028,  $p < .001$ ). The results were expected, as XGB, which uses gradient loss correction with penalization, is more advanced than the other machine learning techniques. Notably, the authors did not employ deep learning methods.

In a multi-center validation study, Lee et al. (2022) developed a machine learning model for preoperative prediction of postoperative mortality, using only 12–18 clinical variables for model training. The data came from 454,404 patients aged 18 or older who underwent non-cardiac surgeries across four independent institutions. The study compared the predictive performances of logistic regression, random forest, extreme gradient boosting (XGBoost), and deep neural network methods. To enhance model robustness and prevent overfitting, they employed bootstrapping and grid search with tenfold cross-validation. XGBoost achieved the best performance, with an AUROC of 0.9376 and an area under the precision-recall curve (AUPRC) of 0.1593. Despite the high AUROC, the study reported a relatively low AUPRC value.

#### **2.1.1.4 Support Vector Machines**

The Support Vector Machine (SVM) is designed to solve binary classification problems in supervised learning, where a set of examples  $X_n \in RD$  is paired with corresponding binary labels  $y_n \in \{+1, -1\}$ . Given a training dataset consisting of example-label pairs  $\{(x_1, y_1), \dots, (x_N, y_N)\}$ , the algorithm estimates model parameters to minimize classification errors (Deisenroth, Faisal, and Ong, 2020). The SVM seeks to identify a set of model parameters that effectively minimize classification errors. This process involves

constructing a decision boundary that maximally separates examples with different labels. SVM is known for delivering high accuracy and generalization across many applications, from text categorization to image classification, with strong theoretical performance guarantees (Steinwart and Christmann, 2008).

A distinctive aspect of SVM is its reliance on geometric principles rather than probabilistic ones. In contrast to the maximum likelihood estimation (MLE) approach, where a probabilistic model is formulated based on an assumed distribution of the data, SVM operates on a **geometric perspective**. This perspective involves building a specific optimization function informed by geometric insights, such as the concepts of inner products, margins, and projections. The SVM approach focuses on finding a hyperplane or decision boundary that best separates the data points according to their labels. By maximizing the margin between the classes, SVM effectively reduces the risk of classification errors and improves model robustness to unseen data.

In typical machine learning models, such as those using MLE or Bayesian inference, the data is represented through probabilistic models. Here, the model's structure and parameters are inferred to reflect the likelihood of the observed data. Conversely, SVM views the problem through the lens of **empirical risk minimization (ERM)**, aiming to minimize classification errors based on actual observed data. The optimization function within SVM is explicitly designed to increase the separation between different classes within the  $RD$  space. This is accomplished by adjusting the weights of the model to enhance the margin between positive and negative samples, ensuring that examples sharing the same label are clustered in the same region, and distinct from other labels. This

geometric orientation sets SVM apart from probabilistic and Bayesian methods, as it concentrates solely on spatial separation and data geometry to achieve high classification accuracy (Deisenroth, Faisal, & Ong, 2020).

It uses concepts like inner products and projections, focusing on optimizing a specific function during training based on geometric intuitions (Deisenroth, Faisal, and Ong, 2020).

In essence, many classification algorithms aim to represent data in  $\mathbf{R}^D$  and partition the space so that examples with the same label occupy the same region, ensuring separation from other examples.

The algorithm considers a convenient partition by linearly splitting the data space into two halves using a hyperplane. Given a data point  $x \in \mathbf{R}^D$ , the goal of SVM in a two-class learning task is to identify the optimal classification function to distinguish between the two classes in the training data. The concept of the "best" classification function can be understood geometrically. For a linearly separable dataset, this corresponds to a separating hyperplane  $f(x)$ , which lies between the two classes. Once this function is determined, a new data point  $x_n$  can be classified by checking the sign of  $f(x_n)$ ;  $x_n$  is assigned to the positive class if  $f(x_n) > 0$ . Since there are multiple potential hyperplanes, SVM aims to identify the best one by maximizing the margin between the two classes. The margin refers to the amount of separation between the classes as defined by the hyperplane. Geometrically, the margin is the shortest distance between the closest data points and the hyperplane. This approach ensures that only a few hyperplanes qualify as the solution for SVM, even though many could exist. The data points closest to the hyperplane, which has the minimum distance to it, are known as the support vectors (Witten and Frank, 2005).

### 2.1.1.5 Kernels

The modular nature of SVM allows for flexibility by treating the choice of classification method (SVM) and the feature representation  $\phi(x)$  as independent decisions. This separation enables the exploration of both aspects individually. Since  $\phi(x)$  can be a nonlinear function, SVM, which typically assumes a linear classifier, can be adapted to create classifiers that behave nonlinearly concerning the input data  $x_n$ . For certain types of similarity functions known as kernels, the function implicitly defines a nonlinear feature transformation  $\phi(x)$ . A kernel is a function  $k: X \times X \rightarrow \mathbb{R}$ , for which there exists a Hilbert space  $H$  and a feature mapping  $\phi: X \rightarrow H$  such that the kernel can be expressed as  $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ . Each kernel  $k$  is uniquely associated with a reproducing kernel Hilbert space (RKHS) (Berlinet and Thomas, 2011). In this association, the canonical feature map is defined as  $\phi(x) = k(\cdot, x)$ . The "kernel trick" (Schölkopf, Smola, and Bach, 2002; Shawe-Taylor and Cristianini, 2004) generalizes the inner product to a kernel function, effectively concealing the explicit nonlinear feature transformation. The matrix  $K \in \mathbb{R}^{N \times N}$ , derived from applying  $k(\cdot, \cdot)$  to a dataset or through inner products, is referred to as the Gram matrix or kernel matrix. For a function to serve as a kernel, it must be symmetric and positive semi-definite, ensuring that any kernel matrix  $K$  is symmetric and satisfies  $\forall z \in \mathbb{R}^N: z^T K z \geq 0$ . There is a unique reproducing kernel Hilbert space associated with every kernel  $k$  (Berlinet and Thomas 2011). In this unique association,  $\phi(x) = k(\cdot, x)$  is called the canonical feature map. The generalization from an inner product to a kernel function is known as the kernel trick (Schölkopf, Smola and Bach, 2002; Shawe, Cristianini, 2004), as it hides away the explicit non-linear feature map. The matrix  $K \in \mathbb{R}$ :

$N \times N$ , resulting from the inner products or the application of  $k(\cdot, \cdot)$  to a dataset, is called the Gram matrix, and is often just referred to as the kernel matrix. Kernels must be symmetric and positive semidefinite functions so that every kernel matrix  $K$  is symmetric and positive semidefinite  $\forall z \in \mathbb{R}^N : z^T K z \geq 0$ . Some commonly used kernels for multivariate real-valued data  $x_i \in \mathbb{R}^D$  include the polynomial kernel, the Gaussian radial basis function (RBF) kernel, and the rational quadratic kernel (Schölkopf, Smola, and Bach, 2002; Rasmussen and Williams, 2006). The primary tuning parameter in SVMs is the cost penalty, which determines how heavily misclassified observations are penalized—higher values impose stricter penalties. The modular nature of Support Vector Machines (SVMs) provides flexibility by separating the choice of classification method and feature representation. This separation allows researchers to independently explore the classification technique (SVM) and the feature representation  $\phi(x)$ , resulting in more robust experimentation with data. Notably, SVM assumes a linear classifier by default, but can be adapted to nonlinear data using an appropriate nonlinear transformation of the input features. This transformation can be done through a kernel function, which effectively allows SVM to create nonlinear decision boundaries (Schölkopf, Smola, and Bach, 2002; Shawe-Taylor and Cristianini, 2004).

A kernel function,  $k(x, y)$ , is a measure of similarity between data points in a transformed feature space. Essentially, it implicitly defines the nonlinear transformation  $\phi(x)$  without requiring the explicit calculation of feature vectors. The beauty of the kernel trick is its ability to avoid the computational complexity of transforming data into high-dimensional spaces. For instance, the polynomial kernel, Gaussian radial basis function (RBF) kernel,



and rational quadratic kernel are widely used in practice (Berlinet and Thomas-Agnan, 2011). These kernels provide SVM with the power to classify data with complex, nonlinear patterns, making SVMs a competitive choice even compared to deep learning models, especially in small or moderately-sized datasets.

However, despite their success, SVMs also have limitations. As the complexity of the data grows—both in terms of features and the number of observations—the computational cost of building and training an SVM model increases dramatically. Large-scale datasets, often found in modern applications, can significantly slow down SVM training times. Moreover, as Huang et al. (2016) demonstrated, while SVMs perform remarkably well on relatively small and structured datasets like those in medical applications, they can struggle when applied to highly unstructured or massive data sources.

In a study by Huang et al. (2016), a support vector machine (SVM) model was developed to predict mortality in burn patients and compared with a logistic regression (LR) model. The overall mortality in this study was 1.8%. Univariate associations with mortality were identified, and independent associations were determined through multivariate logistic regression analysis. Factors independently associated with mortality at admission included gender, age, total burn area, full-thickness burn area, inhalation injury, shock, time to admission, and others.

The logistic regression model demonstrated a sensitivity of 99.75%, a specificity of 85.84%, and an area under the receiver operating characteristic (AUROC) of 0.989 (95% CI: 0.979–1.000;  $p < 0.01$ ). The model correctly classified 99.50% of cases. The subsequently developed SVM model exhibited even better performance, correctly

classifying nearly 100% of the test cases. It was also robust in predicting mortality for both adult and paediatric patients, with accuracy ranging from 92% to 100%. The study did not employ deep learning techniques.

Wallert et al. (2017) developed a machine-learning model to predict two-year survival versus non-survival following a first myocardial infarction (MI). The study used data from 51,943 first MI cases registered over six years (2006–2011) in the Swedish national quality registry, SWEDHEART/RIKS-HIA, which covers 90% of all MIs in Sweden, with follow-up data obtained from the Cause of Death register (over 99% coverage). A Support Vector Machine (SVM) with a radial basis function kernel, using 39 predictors, achieved the best performance on the test set with an area under the receiver operating characteristic curve (AUROC) of 0.845, a positive predictive value (PPV) of 0.280, and a negative predictive value (NPV) of 0.966. The SVM model outperformed the Boosted C5.0 model (AUROC: 0.845 vs. 0.841,  $P = 0.028$ ), though its performance was not significantly better than Logistic Regression or Random Forest. As the sample size and number of predictors increased, the models began to converge, showing minimal differences in performance across algorithms. Given the similarity of results across four machine learning techniques, the authors suggested that deep learning would be a logical next step, particularly with the large and complex dataset. The low PPV of 0.28 indicates room for improvement in the model's predictive accuracy.

#### **2.1.1.6 Deep learning methods**

A deep neural network (DNN) is an advanced structure of interconnected artificial neurons, known as nodes, designed to emulate the neural pathways found in the human brain

(Goodfellow, Bengio, and Courville, 2016). Each of these nodes includes several components: an input, a weight, a bias, and an activation function, which together process inputs to generate an output (Chollet, 2018). This basic unit within a neural network is referred to as a perceptron, and when these perceptrons are connected, they form a network layer. A typical DNN consists of multiple layers, including an input layer, one or more hidden layers, and an output layer, through which data passes sequentially (Choi et al., 2020). As the network grows deeper with additional hidden layers, it gains the ability to capture increasingly complex patterns, which makes it particularly effective for intricate tasks like image recognition, natural language processing, and more.

Deep learning has revolutionized fields such as speech recognition, image classification, and natural language processing by enabling machines to learn complex data patterns. This success stems from the use of multi-layer neural networks (DNNs) that progressively extract more abstract representations of data through layers of transformation. The hierarchical learning approach of DNNs allows simple features to combine into more complex ones, and while the concept may be straightforward, it has been remarkably effective in tasks such as autonomous driving and speech recognition (LeCun, Bengio, & Hinton, 2015). DNNs are structured into three main layers: the input layer, hidden layers, and the output layer. The input layer takes the original data features, and the hidden layers process these features by applying non-linear transformations to uncover meaningful patterns. This ability to learn abstract data representations makes DNNs highly effective at predicting outcomes in complex tasks (Bengio, Courville, & Vincent, 2013).

However, deep learning models often require large datasets and significant computational resources. Historically, such requirements rendered deep learning impractical, but advances in hardware—such as more powerful CPUs and GPUs—along with the increasing availability of data, have allowed deeper and more complex networks to be developed. These advancements have enabled DNNs to identify intricate patterns within large datasets, something previously unattainable due to hardware constraints.

Despite their successes, deep neural networks (DNNs) encounter notable limitations, particularly in the realms of uncertainty quantification and overfitting, especially when data is limited. Overfitting manifests when models excel on training datasets yet struggle to generalize to unseen data. This issue becomes more pronounced as deep learning models often produce overly confident predictions, even in the presence of uncertainty within the data (Ghahramani, 2016).

To mitigate these challenges, probabilistic deep learning models, such as Bayesian Neural Networks (BNNs), present a promising alternative. These probabilistic models effectively represent the uncertain elements of an experiment through probability distributions, offering a cohesive set of tools from probability theory for tasks related to modeling, inference, prediction, and model selection. Central to probabilistic modeling is the joint distribution  $p(x;\theta)$  of the observed variables  $x$  and the hidden parameters  $\theta$ . This distribution encapsulates crucial information derived from the prior and the likelihood, reflecting the product rule. Additionally, the marginal likelihood  $p(x)$ , which is pivotal for model selection, can be computed by integrating out the parameters, following the sum rule. The posterior distribution, obtained by dividing the joint by the marginal likelihood,

is unique to the joint distribution itself. Consequently, a probabilistic model is defined by the joint distribution of all its random variables.

In contrast to traditional neural networks that depend on fixed weight estimates, BNNs utilize probability distribution to capture uncertainty in their predictions (Kendall & Gal, 2017). Within the framework of Bayesian inference, a global probability model is constructed by integrating two fundamental components (Dempster, 1968). The first component, known as the prior distribution, embodies initial beliefs about the parameters before observing any data, reflecting assumptions about their likely values based on existing information or theoretical insights. The second component, the likelihood function, illustrates the relationship between the parameters and the observed data, indicating how probable the observed data is for various parameter values. Collectively, the prior and likelihood facilitate updated inferences about the parameters in light of new observations (Dempster, 1968).

In Bayesian analysis, inferences are articulated as probabilities that quantify the likelihood of specific outcomes or parameter values, particularly when certain parameters remain unknown or when observations are yet to occur. These probabilities integrate prior knowledge encoded in the model with data-driven updates derived from observed information (Dempster, 1968). Essentially, Bayesian inference merges prior beliefs with empirical evidence from data to establish a probabilistic framework for understanding parameter values and making predictions about unseen data. This framework adapts dynamically as new information becomes available. Consequently, BNNs are better equipped to generalize effectively, especially in scenarios involving limited datasets, by

incorporating priors and regularizing model complexity. This adaptability enables BNNs to provide more reliable predictions by adequately accounting for uncertainty (Sun et al., 2019).

Flipout is a significant development in the training of Bayesian neural networks (BNNs) through variational inference. Aimed at improving model efficiency and robustness, Flipout minimizes the variance in gradient estimates by decorrelating them during backpropagation, allowing for faster and more stable convergence (Wen et al., 2018). Unlike conventional regularization techniques that focus on network activations, Flipout's unique approach works by creating pseudo-independent weight perturbations across mini-batches, enabling more efficient utilization of computational resources. Traditional weight perturbation methods are computationally intensive because they often require individual calculations for each sample, limiting their scalability. Flipout overcomes this by leveraging a distribution factorized by weight and centered symmetrically around zero, thus supporting various network architectures such as fully connected networks, convolutional neural networks, and recurrent neural networks (Wen et al., 2018).

The main advantage of Flipout lies in its ability to decorrelate gradient updates without adding bias, which improves the overall gradient quality and speeds up convergence. By generating decorrelated random perturbations in weight parameters, Flipout makes BNNs more robust and enhances their ability to quantify uncertainty. This is particularly valuable in high-stakes areas like healthcare, autonomous driving, and financial forecasting, where models must not only predict accurately but also provide well-calibrated confidence levels.

In these applications, reliable uncertainty estimation is essential to support decision-making, as BNNs combined with Flipout can help prevent overfitting, a common issue in deep neural networks, especially when working with limited data (Wen et al., 2018; Joshi & Dhar, 2022).

Moreover, Flipout contributes to the accuracy and calibration of BNN predictions by generating weight distributions rather than single-point estimates. This allows models to capture a broader range of data variability and deliver predictions with more accurate confidence estimates. This combination of regularization and enhanced uncertainty handling aligns the model's output probabilities more closely with actual likelihoods, helping create better-calibrated predictions. Ultimately, Flipout's ability to balance computational efficiency with robustness makes it an ideal tool for probabilistic modeling in deep learning (Wen et al., 2018).

Moreover, BNNs help control model complexity, a key factor in preventing overfitting. Integrating prior knowledge into the learning process allows BNNs to automatically manage model complexity, improving generalization, even with limited data. Over time, as more data becomes available, BNNs can make increasingly deterministic predictions, thus enhancing their performance in various applications (Blundell et al., 2015).

The probabilistic nature of BNNs, coupled with the efficiency gains from Flipout, makes these models particularly suitable for applications where uncertainty quantification is vital, such as medical diagnosis and autonomous driving. While traditional neural networks are primarily concerned with improving accuracy, BNNs, enhanced by Flipout, address both

accuracy and uncertainty, thus reducing the risk of overfitting and improving performance in uncertain environments (Srivastava et al., 2014; Wen et al., 2018). This combination has also been shown to outperform traditional dropout techniques, introducing less noise into the training process and enabling the model to retain more accurate feature representations. In addressing the problem of data scarcity, BNNs and Flipout can also be combined with data augmentation techniques to enhance model generalization. Generative models, such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), can further complement BNNs by generating synthetic data, thus mitigating the risk of overfitting in scenarios with limited data.

In conclusion, while deep learning models such as DNNs have demonstrated exceptional performance across various tasks, their limitations in uncertainty quantification, data efficiency, and overfitting necessitate the adoption of more advanced approaches like BNNs and Flipout. The integration of probabilistic deep learning methods, particularly BNNs enhanced by Flipout, represents a crucial step toward improving the reliability and robustness of models in real-world applications requiring both high accuracy and effective uncertainty management.

Nudel et al. (2021) conducted a comparison between two machine learning techniques—artificial neural networks (ANNs) and gradient boosting machines (XGBs)—against traditional logistic regression (LR) models in predicting anastomotic leaks and venous thromboembolism (VTE) following bariatric surgery. The models were trained and validated using preoperative data from 2015-2017, with a study cohort of 436,807 patients. The incidences of leaks and VTE were 0.70% and 0.46%, respectively. For predicting



leaks, ANN proved to be the most effective, achieving an AUC of 0.75 (95% CI, 0.73-0.78), followed by XGB with an AUC of 0.70 (95% CI, 0.68-0.72), and LR trailing with an AUC of 0.63 (95% CI, 0.61-0.65), with all comparisons showing statistical significance ( $p < 0.001$ ). For VTE detection, the performance of ANN, XGB, and LR was comparable, with AUCs of 0.65 (95% CI, 0.63-0.68), 0.67 (95% CI, 0.64-0.70), and 0.64 (95% CI, 0.61-0.66), respectively, although the difference between XGB and LR was statistically significant ( $p = 0.001$ ). Both ANN and XGB models surpassed traditional LR in predicting anastomotic leaks. Kasim et al. (2022) developed models to predict mortality among elderly patients presenting with ST-elevation myocardial infarction (STEMI) using both deep learning (DL) and traditional machine learning (ML) algorithms, including logistic regression (LR), random forests (RF), XGBoost (XGB), and support vector machines (SVM). These models were compared with common scoring systems like Thrombolysis in Myocardial Infarction (TIMI). Their study focused on integrating DL with ML feature selection techniques to better understand DL's "black box" nature, particularly in predicting mortality among elderly STEMI patients in an Asian cohort. The researchers hypothesized that combining DL with ML-based feature selection algorithms would improve in-hospital mortality predictions for this group.

The main performance metric was the area under the receiver operating characteristic curve (AUC). The features selected by RF, XGB, SVM, and LR were applied to the DL model. The results indicated that ML models built using a reduced set of features outperformed those developed with the full set. For example, the AUCs for LR (0.91 vs. 0.83), RF (0.91 vs. 0.89), XGB (0.89 vs. 0.89), and SVM (0.91 vs. 0.87) were higher when fewer features

were used. The DL model using all features achieved an AUC of 0.93, slightly better than the ML models using reduced feature sets. However, the DL model constructed with selected features (e.g., AUC of 0.95 using RF-selected variables) demonstrated superior performance compared to the model using all features (AUC 0.93). There was no statistically significant difference between the DL models developed with selected features from various ML algorithms ( $p > 0.05$ ).

In terms of positive predictive value (PPV), the models varied: LR (0.34), RF (0.57), SVM (0.53), XGB (0.49), and DL (0.43). While these metrics were reported, the authors did not provide comparisons based on the F1 score. They also highlighted a key limitation of DL—that it lacks built-in feature importance mechanisms. Unlike ML models, which provide clear feature rankings, DL models automatically learn patterns from the data, making feature importance less transparent. To address this, the authors applied features selected from ML models (RF, XGB, SVM, and LR) to develop the DL models. Although DL using all features had slightly lower AUCs (0.93) than DL models using selected features (e.g., AUC 0.95 with RF-selected variables), there was no statistically significant difference between the models.

The approach of developing a DNN with selected features contradicts traditional deep learning principles, where DNNs are typically designed to extract relevant features autonomously from raw data (Cosgriff & Celi, 2020). However, feature importance in DL models can be derived using interpretability techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), which offer insights into model decision-making.

Nilsson et al. (2006) developed artificial neural networks (ANNs) to predict mortality based on preoperative evaluations from the EuroSCORE database, which included 18,362 patients. Out of the original 72 risk variables, the model identified 34 as relevant for mortality prediction. The ANN model demonstrated a higher area under the receiver operating characteristic curve (AUC) of 0.81 compared to the logistic regression-based European System for Cardiac Operative Risk Evaluation (EuroSCORE), which had an AUC of 0.79 ( $P < .0001$ ). However, the authors did not report other key performance metrics, such as precision, recall, or F1 score, which makes it difficult to fully assess the model's effectiveness and implications beyond AUC.

Krittanawong et al. (2021) applied machine learning (ML) and deep learning (DL) techniques to predict mortality in patients with spontaneous coronary artery dissection, a rare cause of acute coronary syndrome (ACS), using a relatively small dataset of 375 patients. The ML models tested included logistic regression, support vector machine (SVM), decision tree, random forest, K-nearest neighbors, AdaBoost, and extreme gradient boosting (XGBoost). The DL model employed a deep neural network built with Python (Keras and TensorFlow backend). To address class imbalance in the event data, random over-sampling was utilized.

The models were evaluated based on the area under the receiver-operator characteristic curve (AUC) and adjusted for class imbalance. The DL model was the best performer, achieving an AUC of 0.98 (95% CI 0.97–0.99), accuracy of 98%, sensitivity of 98%, and specificity of 96%. In comparison, the AdaBoost model had an AUC of 0.95 (95% CI 0.93–0.96), accuracy of 61%, sensitivity of 25%, and specificity of 97%. The SVM model

yielded an AUC of 0.92 (95% CI 0.89–0.94), with accuracy at 60%, sensitivity at 25%, and specificity at 96%. Other ML models also performed differently: K-nearest neighbors achieved an AUC of 0.91 (95% CI 0.88–0.93), accuracy of 50%, sensitivity of 74%, and specificity of 97%. XGBoost had an AUC of 0.90 (95% CI 0.86–0.93), accuracy of 54%, sensitivity of 83%, and specificity of 99%. The decision tree model had an AUC of 0.78 (95% CI 0.72–0.83), accuracy of 53%, sensitivity of 87%, and specificity of 35%. Logistic regression performed poorly, with an AUC of 0.59 (95% CI 0.51–0.67), accuracy of 59%, sensitivity of 25%, and specificity of 94%. The random forest model had the lowest AUC of 0.50 (95% CI 0.41–0.58), accuracy of 52%, sensitivity of 25%, and specificity of 96%, showing no significant difference from logistic regression.

The authors attributed the DL model's superior performance to its ability to handle multidimensional variables, potentially capturing more complex interactions through matrix multiplication, weights, and biases than kernel methods and regularization penalties used in SVM. However, they did not employ explainability techniques like LIME or SHAP to identify the features that contributed to the model's predictions, leaving the interpretability of the DL model unclear.

Chen et al. (2022) developed a model to predict 30-day postoperative mortality by incorporating deep neural networks (DNN) and natural language processing (NLP) techniques, specifically using BERT (Bidirectional Encoder Representations from Transformers). Their innovative approach included unstructured clinical text data (such as preoperative diagnoses and proposed procedures) to enhance postoperative mortality prediction. The BERT-DNN model achieved the highest area under the receiver operating

characteristic curve (AUROC) of 0.964 (95% CI 0.961-0.967) and an area under the precision-recall curve (AUPRC) of 0.336 (95% CI 0.276-0.402). The BERT-DNN model outperformed traditional logistic regression (AUROC = 0.952, 95% CI 0.949-0.955) and the American Society of Anesthesiologists Physical Status classification (ASAPS; AUROC = 0.892, 95% CI 0.887-0.896). However, the difference in AUROC between BERT-DNN and other models such as the standard DNN (AUROC = 0.959, 95% CI 0.956-0.962) and random forest (AUROC = 0.961, 95% CI 0.958-0.964) was not statistically significant.

In terms of AUPRC, the BERT-DNN model showed a significant improvement over all other models, including the DNN (AUPRC = 0.319, 95% CI 0.260-0.384), random forest (AUPRC = 0.296, 95% CI 0.239-0.360), logistic regression (AUPRC = 0.276, 95% CI 0.220-0.339), and ASAPS (AUPRC = 0.149, 95% CI 0.107-0.203). Despite these promising results, the authors did not incorporate explainability techniques such as SHAP or LIME, leaving the interpretation of feature importance or insights into the decision-making process of the model unaddressed.

Ahmed et al. (2020) developed a deep neural network (DNN) model to predict mortality in trauma patients admitted to the intensive care unit. Their approach involved identifying statistically significant risk factors from the dataset, which were then input into the DNN model for mortality prediction. The DNN model demonstrated strong performance, with a training accuracy of 93.8% and a testing accuracy of 92.3%. Additional metrics included a sensitivity of 79.1%, specificity of 94.2%, positive predictive value (PPV) of 66.42%, negative predictive value (NPV) of 96.87%, and an area under the receiver operating

characteristic curve (AUROC) of 0.91. When compared to other conventional machine learning models—such as Linear Discriminant Analysis (LDA), Gaussian Naive Bayes (GNB), K-Nearest Neighbors (KNN), and Decision Tree (CART)—the DNN outperformed them in predictive power. However, a critical limitation of the study is the lack of model interpretability. While the model delivers excellent predictive results, the absence of techniques like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) makes it difficult to understand which features are driving these predictions. This issue is particularly relevant in medical decision-making, where understanding the "why" behind a prediction is crucial for clinical acceptance and trust. Without explainability, the model remains a "black box," making it harder for clinicians to justify relying on it for life-or-death decisions.

Lee et al. (2018) developed a deep neural network (DNN) aimed at predicting postoperative in-hospital mortality using 87 intraoperative features. These features, selected through clinical consensus, included vital signs like minimum and maximum blood pressure, interventions such as the total amount of blood and fluids administered, and anesthesia-related descriptions such as the presence of an arterial line and type of anesthesia. This diverse set of features allowed the model to capture a broad range of intraoperative conditions, potentially predictive of patient mortality following surgery. To evaluate which features had the most predictive power within the model, the authors performed a **feature ablation analysis**. This method involves removing groups of features, retraining the model with the same architecture and hyperparameters, and then assessing the model's performance changes—specifically, changes in the area under the receiver operating

characteristics curve (AUROC). By observing how the removal of specific features affected performance, they aimed to quantify the importance of different feature sets to overall model accuracy. The deep neural network consistently outperformed logistic regression models in most feature combinations, demonstrating that the DNN captured complex, non-linear relationships in the data better than traditional methods. Interestingly, even when the feature set was reduced from 87 to 45 variables, the DNN's performance remained high, indicating that the model was able to generalize well even with fewer inputs. Adding preoperative mortality scores and ASA (American Society of Anesthesiologists) scores further improved model performance, especially in the reduced feature set. Despite the model's predictive strength, the study has several limitations. First, the use of intraoperative features—some specific to complex surgeries—limits its generalizability to a broader patient population. As noted by the authors, the model may not be directly applicable to other types of surgeries or clinical scenarios. This complexity makes the model less accessible for widespread use in various surgical settings. Another critical limitation lies in the method used to assess feature importance. The authors employed feature ablation analysis to evaluate the effect of removing features on the model's performance. While this method provides some insight, it is less sophisticated compared to modern techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive explanations), which offer more granular, interpretable explanations about the contribution of individual features. These modern techniques can better address the black-box nature of deep neural networks by providing detailed explanations of how and why certain features impact.

### **2.1.1.7 The Challenge of Overfitting in SVM and Deep Learning**

Overfitting is a challenge that both SVM and deep learning models face. In the case of SVMs, the primary control mechanism for overfitting is the cost parameter, which regulates the trade-off between the misclassification of training examples and the simplicity of the decision boundary. A higher cost results in a stricter penalty for misclassified observations, thus leading to a more complex model that is more likely to overfit, especially in noisy datasets (Schölkopf, Smola, and Bach, 2002). Deep learning models, particularly deep neural networks (DNNs), also struggle with overfitting, especially when trained on limited data. This is due to their highly flexible and parameterized structure, which allows them to model even minute details in the training data, including noise (Ghahramani, 2016).

Overfitting is further exacerbated in DNNs due to the large number of parameters involved. DNNs rely on large datasets to prevent overfitting and generalize well to unseen data. Probabilistic approaches such as Bayesian Neural Networks (BNNs) attempt to address this by incorporating prior knowledge about the model parameters and treating them as distributions rather than fixed values (Sun et al., 2019). BNNs help control overfitting by incorporating uncertainty into the model. This uncertainty is captured by learning distributions over the weights instead of point estimates. In doing so, BNNs provide not only point predictions but also a measure of uncertainty in their predictions. This is particularly useful when working with limited data or when encountering noisy or ambiguous inputs (Wang and Yeung, 2016).



### **2.1.1.8 Applications of SVMs, Deep Learning, and BNNs**

SVMs, deep learning models, and BNNs have been applied to a wide range of fields, from image recognition to healthcare diagnostics. For example, Huang et al. (2016) used an SVM model to predict mortality in burn patients, achieving near-perfect accuracy. Similarly, Wallert et al. (2017) applied an SVM with a radial basis function kernel to predict two-year survival after myocardial infarction, demonstrating that traditional machine learning models like SVM can still be highly competitive.

In the context of deep learning, recent studies have shown that incorporating probabilistic models, such as BNNs, can significantly enhance the reliability and interpretability of predictions. This is especially crucial in sensitive fields such as healthcare, where understanding the uncertainty of a model's prediction can be just as important as the prediction itself. Additionally, by utilizing the Flipout estimator, models can improve their robustness and convergence speed, leading to more accurate and reliable predictions, even with smaller datasets.

Overall, while deep learning and SVMs have become mainstays in modern machine learning applications, the integration of probabilistic techniques like BNNs and Flipout is paving the way for more reliable and uncertainty-aware models. These advancements are particularly valuable as machine learning continues to tackle increasingly complex and high-risk tasks. The challenges faced by both SVMs and DNNs, particularly with respect to overfitting and uncertainty quantification, highlight the importance of probabilistic approaches like BNNs. By incorporating uncertainty directly into the model's structure and leveraging techniques like Flipout, modern deep learning models can achieve better

generalization, especially in low-data regimes. Future research should continue exploring the combination of traditional machine learning methods, like SVMs, with probabilistic deep learning techniques to further enhance model robustness and accuracy in various application domains.

### **2.1.2 Explainable AI Methods - Overview**

Artificial Intelligence (AI) is now foundational in numerous sectors that have embraced modern information technologies (Russell & Norvig, 2016). Although the origins of AI span several decades, there is a strong consensus on the critical role of intelligent systems equipped with capabilities like learning, reasoning, and adaptation. These capabilities enable AI to achieve unprecedented results across increasingly complex computational tasks, cementing its importance for future societal progress (West, 2018). Recent advancements in AI-driven systems have reached a level where minimal human input is required for their design and deployment. However, in contexts where AI-derived decisions impact human lives—such as in healthcare, law, or defense—there is an emerging need for transparency around the decision-making process of these systems (Goodman & Flaxman, 2017).

In the early stages, AI systems were more interpretable and straightforward to understand. Yet, recent years have seen the rise of opaque decision-making models, particularly in the form of Deep Neural Networks (DNNs). The empirical success of Deep Learning (DL) models like DNNs is attributed to both efficient learning algorithms and their extensive parametric space. This vast parameter count, consisting of hundreds of layers and millions of individual parameters, leads DNNs to be regarded as complex black-box models

(Castelvecchi, 2016). In contrast to this black-box nature is the concept of transparency, which emphasizes the importance of clear insights into the mechanisms behind a model's function (Lipton, 2018). Deep learning has significantly revitalized machine learning research by showcasing its ability to learn from vast datasets to solve complex problems, often achieving performance levels that surpass humans in certain tasks (Mnih et al., 2015). This success has propelled AI into the mainstream, but its strengths also present challenges. Deep learning models, which consist of millions of parameters, are highly complex (Hu et al., 2021), making them difficult for humans to interpret (Lakkaraju, Arsov, and Bastani, 2020). As these "black-box" models are increasingly used in critical, high-stakes areas like medical AI and autonomous driving, the consequences of their failures become more serious, such as medical errors or accidents involving autonomous vehicles. To address this, model-agnostic methods for explaining machine learning predictions have gained prominence. These techniques treat models as black-box functions, offering greater flexibility in selecting models, explanations, and representations. This enhances capabilities in debugging, model comparison, and creating interfaces tailored for different users and models (Ribeiro, Singh, and Guestrin, 2016). One of the primary challenges in fostering a shared understanding in the AI field is the frequent misuse and conflation of the terms interpretability and explainability in the literature. These concepts, while related, embody distinct characteristics. Interpretability refers to a model's passive quality or the extent to which a model appears logical to a human observer—this quality is often referred to as *transparency* (Lipton, 2018). On the other hand, explainability is an active quality,

involving any actions or processes that clarify or reveal a model's internal mechanics for human users (Guidotti et al., 2018).

To clarify commonly used terms in ethical AI and Explainable AI (XAI) discourse, the terms are summarized as under with key distinctions and similarities in nomenclature:

- Understandability (also called intelligibility) refers to a model's capacity to allow a human to understand its functionality—how it works—without necessarily detailing its internal workings or algorithms (Mueller et al., 2019).
- Comprehensibility for machine learning models indicates the capacity of a learning algorithm to express its learned knowledge in a way understandable to humans. This concept, derived from Michalski's postulates, suggests that “the outcomes of computer induction should resemble symbolic descriptions that an expert might produce when observing the same data. These descriptions should represent coherent information chunks, interpretable in natural language and combining quantitative and qualitative insights” (Michalski, as cited in Guidotti et al., 2018).
- Interpretability is defined as the ability to elucidate or provide meaning in understandable terms for human users.
- Explainability is conceived as an interface between humans and an AI decision-maker, offering both an accurate representation of the decision-maker's logic and a format that is clear to human observers (Guidotti et al., 2018).
- Transparency pertains to a model's inherent understandability. Transparent models can be classified into simulatable models, decomposable models, and

algorithmically transparent models, each differing in how they facilitate user understanding of the model's operations (Lipton, 2018).

Across these definitions, understandability is fundamental in XAI. Both transparency and interpretability closely relate to understandability: while transparency involves a model's inherent capacity to be understood, interpretability represents the ease with which a human can comprehend a model's decisions. Similarly, comprehensibility ties into understandability by emphasizing the audience's ability to grasp the knowledge embedded in the model. Thus, understandability involves both the model's clarity and the human observer's capacity to interpret the model's output. When developing a machine learning (ML) model, considering interpretability as an additional design factor can significantly enhance the model's implementability for three key reasons:

1. **Impartial Decision-Making:** Interpretability can help ensure fair decisions by enabling the identification and correction of biases in the training dataset (Hall, 2018).
2. **Robustness against Adversarial Perturbations:** A clearer understanding of the model's decision boundaries aids in identifying areas susceptible to adversarial manipulation, thereby improving robustness (Hall, 2018).
3. **Meaningful Variable Influence:** By focusing on interpretable models, we can confirm that only relevant variables contribute to the output, suggesting the existence of a genuine causal relationship within the model's reasoning (Hall, 2018).

For these reasons, an interpretable ML system should ideally provide insights into the model's mechanisms and predictions, present visualizations of its decision rules, or identify potential perturbations that could influence its outcomes.

Interpretable machine learning models empower healthcare practitioners to make well-informed, data-driven decisions, which can enhance personalized care and overall service quality. Broadly, interpretability methods for ML models are classified into global and local approaches. Traditionally, ML research has emphasized global interpretability, which helps users understand how the model's inputs relate to the entire range of predictions it makes (Bratko, 1997; Martens et al., 2008). In contrast, local interpretability aims to clarify how the model arrives at specific predictions for individual cases or for narrowly defined areas within the prediction space (Hall et al., 2017; Stiglic et al., 2020). Both approaches are critical in different contexts, depending on whether the focus is on the model's general behavior or individual predictions. In the context of Interpretability and explainability.

### **2.1.3 Model-Agnostic Interpretability of Machine Learning**

The concept behind model-agnostic explanation methods is that they treat machine learning models as black-box systems, meaning they rely solely on the model's outputs without needing access to its internal mechanics. This flexibility makes them highly adaptable since these methods do not require details about the model's structure, such as in neural networks, where specifics like topology, weights, biases, or activation values remain unknown (Holzinger et al., 2020). This approach allows model-agnostic methods to be

applied across a broad range of models, enhancing their versatility in different machine-learning contexts.

### **2.1.3.1. LIME (Local Interpretable Model Agnostic Explanations)**

The core concept behind LIME (Local Interpretable Model-agnostic Explanations) is to clarify a prediction made by a complex model, such as a deep neural network (denoted as  $f_M$ ), by creating a simpler, more interpretable model,  $f_S$  called a surrogate model. This surrogate model is easier to understand and explain, which is why LIME is often categorized as a surrogate-based explanation technique (Samek et al., 2021).

LIME aims to create a model that is interpretable and accurately represents the behavior of the original model within the local neighborhood of a specific instance (Ribeiro, Singh, and Guestrin, 2016). Although it may be challenging to replicate the entire black-box model globally with a simpler model, it is feasible to achieve a close approximation in the area around a particular data point. In formal terms, the explanation model is denoted as  $g : R^d \rightarrow R$  where  $g \in G$ , belongs to a set of interpretable models  $G$  (such as linear models, decision trees, or rule lists). This allows the output  $g$  to be shared with the user as a clear, understandable explanation. Since not every model  $g \in G$  is easily interpretable, the complexity of  $g$  noted as  $\Omega(g)$ , acts as a measure to help maintain simplicity. Complexity can be regulated either as a soft constraint (like limiting the depth of a decision tree) or as a hard limit (where the complexity becomes infinitely high if certain thresholds are exceeded).

The black-box model to be explained is  $f: Rd \rightarrow R$ :  $f: Rd \rightarrow Rf$ : which in classification cases represents the probability of a particular class for  $x$ . To assess locality, a proximity measure  $\Pi_x(Z)$  is defined to indicate the closeness between an instance  $z$  and  $x$ . The goal is to measure the dissimilarity,  $L(f, g, \Pi x)$  between  $f$  and  $g$  around the instance  $x$ , ensuring that the explanation model  $g$  both maintains low  $L$  and has manageable complexity  $\Omega(g)$  for interpretability. The final explanation  $\xi(x)$  is determined by solving:

$$\xi(x) = \arg \min_{g \in G} L(f, g, \Pi x) + \Omega(g)$$

This framework supports a variety of explanation families  $G$ , fidelity measures  $L$ , and complexity controls  $\Omega$ . LIME estimates  $L$  by generating variations of the instance  $x$ , predicting them using the black-box model  $f$ , and adjusting the weights according to  $\Pi x$ .

Primarily, The method involves generating a set of samples around the input of interest,  $x_i$ , by exploring its local neighborhood,  $N_{x_i}$ . These samples are then evaluated using the original complex model. LIME then approximates the behaviour of the complex model in this local region by fitting a straightforward linear function, serving as the surrogate model. Essentially, instead of directly explaining the original model's prediction for  $f_M(x_i)$ , LIME provides an explanation based on  $f_S(x_i)$ , where the surrogate model closely mimics the target model in the neighborhood of  $x_i$  (i.e.,  $f_M(x) \approx f_S(x)$  for  $x \in N_{x_i}$ ) (Holzinger et al., 2020). This approach allows for simplifying and interpreting how the complex model behaves in small, localized regions, offering insights without needing to understand the full intricacies of the complex model.



### 2.1.3.2 SHAP (Shapley Values)

The Shapley value method for explainability is grounded in Lloyd Shapley's foundational work in game theory (Shapley, 1953). This approach conceptualizes a regression problem as a cooperative game, where each predictor variable is treated as a "player" in the model's prediction process. This game aims to improve regression accuracy (or equivalently, to reduce error), with the Shapley value quantifying each predictor's unique contribution toward that goal (Molnar, 2022). Determining each predictor's contribution is complex because predictors interact in intricate, often non-linear ways, influenced by the model's structure—such as multiplicative interactions like  $X_1 \cdot X_2$  (Lundberg & Lee, 2017).

Shapley value methods tackle this complexity by evaluating the impact of every possible subset of predictors on the model's output, providing a per-predictor, per-data point estimate of each predictor's influence. This approach offers a comprehensive breakdown of the model's behaviour for individual predictions, adding transparency to complex models (Vowels et al., 2022).

The importance of predictor  $i$  is determined by analysing how adding  $i$  to a subset  $S$  of other predictors affects the function's value  $e_S$ . The contribution of predictor  $i$ , denoted as  $\phi(i)$ , is computed as a weighted average over all possible subsets  $S$  of the other predictors:

$$\phi(i) = \sum_{S \subseteq \{1, \dots, p\} / \{i\}} \frac{|S|! (p - 1 - |S|)!}{p!} (e_{S \cup \{i\}} - e_S).$$

This formula is equivalent to

$$\phi(i) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} e_{\text{before}(\pi, i) \cup \{i\}} - e_{\text{before}(\pi, i)},$$

Here,  $\Pi$  represents the set of all possible orderings of the  $p$  variables, and  $\text{before}(\pi, i)$  refers to the subset of variables that appear before predictor  $i$  in the ordering  $\pi$ . Each ordering describes how the values of  $e_S$  shift as variables are added, starting from  $e_\emptyset$  (the model's baseline) to  $f(x^*)$  (the model's prediction for a specific data point). In essence, the analysis of a single ordering shows how the model output changes as new predictors are added. SHAP values are derived by averaging these contributions across all possible orderings, making the approach robust and fair in terms of allocating credit to each variable.

SHAP (Shapley Additive Explanations) is an extension of the Shapley value framework designed to explain individual predictions of machine learning models. Originally developed to fairly distribute rewards in cooperative games, Shapley values are the only solution that satisfies key fairness principles like efficiency, symmetry, dummy, and additivity. These principles ensure that the contributions are fairly distributed based on each predictor's true impact on the model.

### **2.1.3.3 SHAP (SHapley Additive exPlanation) Values**

SHAP (SHapley Additive exPlanation) values were introduced by Lundberg and Lee (2017) as a unified approach to quantifying feature importance. These values are essentially the Shapley values of a conditional expectation function based on the original model. In this context, they serve as the solution to an equation where  $f_X(z') = f(h_X(z')) = E[f(z) | z_S]$ , and  $S$  represents the set of non-zero indexes in  $z'$ . SHAP values provide a unique, additive method of determining feature importance by utilizing conditional expectations to define simplified inputs.

The concept of SHAP values includes a simplified input mapping, denoted as  $h_{\mathbf{x}}(z')=z^S$ , where  $z^S$  omits the values for features not included in set  $S$ . Since most models are not equipped to handle arbitrary patterns of missing data, the authors propose approximating  $f(z^S)$  by using  $E[f(z)|z^S]$ , the expected value of the model's output given the subset  $S$ . This approach was designed to align closely with other attribution methods, such as Shapley regression, Shapley sampling, and quantitative input influence, while also maintaining connections to techniques like LIME, DeepLIFT, and layer-wise relevance propagation.

By taking this additive and consistent approach, SHAP values offer a comprehensive and flexible method for understanding how individual features contribute to model predictions.

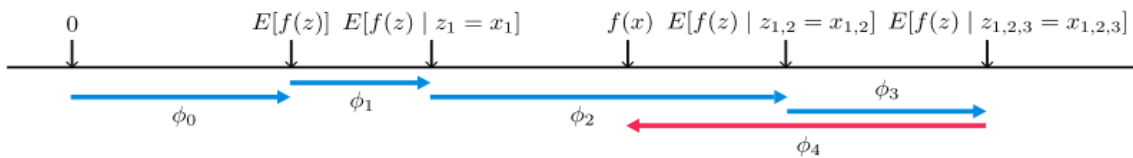


Figure 1: SHAP (SHapley Additive exPlanation) values attribute to each feature the change in the expected model prediction when conditioning on that feature. They explain how to get from the base value  $E[f(z)]$  that would be predicted if we did not know any features to the current output  $f(x)$ . This diagram shows a single ordering. When the model is non-linear or the input features are not independent, however, the order in which features are added to the expectation matters, and the SHAP values arise from averaging the  $\phi_i$  values across all possible orderings.

### Figure 2.1 SHAP (Shapley Additive explanation Values (Lundberg et al.,2017)

Moncada-Torres et al. (2021) conducted a study using data from the Netherlands Cancer Registry, which included 36,658 patients with non-metastatic breast cancer. They compared the performance of Cox Proportional Hazards (CPH) regression with machine learning techniques such as Random Survival Forests, Survival Support Vector Machines,

and Extreme Gradient Boosting (XGB) in predicting survival, with the c-index serving as the performance metric. To interpret the models' predictions, they employed SHAP (Shapley Additive Explanation) values, shedding light on both the classical CPH model and the best-performing machine learning model, XGB.

The SHAP values were used to visualize the overall feature importance within each model. The authors generated SHAP dependence plots, which illustrated how individual features influenced the predictions of each model. By applying SHAP values, they were able to uncover the differences in performance between the reference model (CPH) and the top-performing ML model (XGB), effectively demystifying the "black box" nature of the machine learning model.

Interestingly, the feature importance derived from SHAP values was consistent with the feature significance suggested by p-values in traditional CPH analysis. The study highlighted that SHAP values enabled a detailed examination of how specific features impacted the model's predictions—something that can be challenging even for experts due to the complex and heterogeneous nature of the data. This made SHAP a valuable tool for model interpretation and comparison between traditional and modern approaches.

Lee et al. (2022) developed a predictive model for 30-day mortality following non-cardiac surgery using data from 454,404 patients aged 18 and above, collected from four independent institutions: Seoul National University Hospital (SNUH), AMC, EUMC, and BRMH. The model was trained using a limited set of 12-18 clinical variables. Four different machine learning techniques—logistic regression, random forest, extreme gradient boosting (XGBoost), and deep neural networks—were applied to compare

prediction accuracy. To prevent overfitting and ensure model robustness, the researchers used bootstrapping and grid search with tenfold cross-validation. Among the methods, XGBoost showed the best performance on SNUH data, achieving an AUROC of 0.9376 and an area under the precision-recall curve (AUPRC) of 0.1593. Additionally, when the SNUH model was validated using data from Ewha Womans University Medical Center (EUMC), it achieved an even higher AUROC of 0.941.

SHAP (Shapley Additive Explanation) values were used to determine feature importance across models from each hospital. The results varied across institutions: in SNUH and AMC, albumin emerged as the strongest predictor of 30-day mortality, while in EUMC and BRMH, age and preoperative PT were the most significant predictors, respectively. However, the study was unable to fully explain the differences in feature importance between the institutions, attributing these discrepancies to variations in clinical environments and patient demographics across the hospitals.

Model calibration techniques are essential for adjusting predicted probabilities to better reflect actual outcomes. These methods improve the reliability of probabilistic predictions, which is crucial in clinical settings where accurate probability estimates directly inform decision-making. The most commonly used calibration methods include Platt Scaling, Isotonic Regression, and Temperature Scaling. In the context of mortality prediction, the goal is to ensure that the model's predicted probabilities of mortality closely align with the true outcomes, a task that is complicated by factors such as class imbalance, skewed data distributions, and the high complexity of machine learning models.

Platt Scaling (Platt, 1999) is one of the earliest calibration techniques, commonly applied to models that output uncalibrated scores. This method fits a logistic regression model to the raw output scores of a classifier, effectively transforming them into calibrated probabilities. While it is simple and effective for many models like Support Vector Machines (SVMs), it can underperform with more complex neural networks or models that output highly nonlinear scores.

In contrast, Isotonic Regression (Zadrozny & Elkan, 2002) is a non-parametric calibration technique that can provide a more flexible mapping between the raw output and calibrated probabilities. It is particularly useful for models with more complex output patterns, but it requires large amounts of data to avoid overfitting, making it less suitable for small or highly imbalanced datasets.

Temperature Scaling (Guo et al., 2017) has gained popularity for calibrating Deep Neural Networks (DNNs) and Bayesian Neural Networks (BNNs). This method involves dividing the logits (raw outputs) by a temperature parameter, which softens or sharpens the predicted probabilities. It has shown significant promise in improving calibration without substantial performance loss, especially when the dataset is imbalanced or when dealing with deep learning models, which tend to produce overly confident predictions.

## **2.2 Summary**

### **Consolidated Summary of Literature Review on Artificial Intelligence and Interpretability**

The construction of effective machine learning systems hinges on the interplay of three fundamental components: data, models, and the learning process. This triad forms the

foundation for achieving robust performance, where a model's effectiveness is gauged by its ability to generalize to new, unseen data. Establishing meaningful performance metrics, such as accuracy and error rates, is crucial not only for model improvement but also to ensure predictions closely align with actual outcomes. Automation stands out as a pivotal aspect of machine learning, enabling algorithms to derive insights from data without extensive domain-specific knowledge, thus democratizing access to advanced analytical capabilities. The overarching goal of machine learning is to uncover key patterns within data, allowing models to approximate the mechanisms that generate such data. Well-constructed models capture essential characteristics, enabling accurate predictions about real-world events without necessitating direct experimentation. The machine learning workflow can be delineated into three distinct phases: prediction (or inference), training (or parameter estimation), and hyperparameter tuning (or model selection). During the prediction phase, uncertainty quantification becomes increasingly vital, especially within probabilistic frameworks, as it allows practitioners to express confidence in their predictions. The training phase focuses on optimizing model parameters to enhance performance, often likened to a hill-climbing approach aimed at discovering optimal configurations.

Non-probabilistic models typically employ Empirical Risk Minimization (ERM) to reduce empirical risk but run the risk of overfitting by closely adhering to training data. In contrast, Maximum Likelihood Estimation (MLE) seeks optimal parameters by maximizing the likelihood function without accounting for uncertainty. Probabilistic models utilize Bayesian inference, integrating prior knowledge with likelihoods and posterior

distributions to provide a comprehensive representation of uncertainty. Bayesian methods refine parameter estimates through Maximum A Posteriori (MAP) estimation, effectively merging prior beliefs with observed data.

Regularization techniques are crucial in addressing overfitting in non-probabilistic models by incorporating penalty terms into the loss function, while in probabilistic models, the prior distribution plays a similar role by biasing parameter estimates towards simpler models. Ultimately, non-probabilistic models prioritize optimization without explicit uncertainty modelling, whereas probabilistic models excel in quantifying uncertainty via Bayesian methods. While both MLE and MAP yield point estimates that can obscure valuable information, retaining full posterior distributions is advantageous for enhanced predictive accuracy.

Decision trees, including algorithms like CART, ID3, C4.5, and C5, efficiently partition sample data by splitting variables at specific thresholds, resulting in a structured tree format. CART (Classification and Regression Trees) is particularly versatile in handling both classification and regression tasks, utilizing the Gini index for binary classification and cost-complexity pruning for parameter estimation. Random Forest (RF) amplifies predictive accuracy by aggregating outputs from multiple decision trees built on bootstrapped samples of training data. By employing a random subset of predictors for each tree and using Gini impurity as the splitting criterion, RF not only enhances robustness but also incorporates out-of-bag (OOB) error as an internal validation metric, providing reliable estimates of model performance on unseen data.



Conversely, boosting methods like AdaBoost sequentially train base classifiers, placing more weight on misclassified points to enhance overall model accuracy. Extreme Gradient Boosting (XGBoost) is a powerful gradient boosting implementation that iteratively adds trees to minimize loss while emphasizing parameter adjustments without altering existing trees. Support Vector Machines (SVM) leverage geometric principles to tackle binary classification problems by identifying an optimal decision boundary that maximally separates classes, thus enhancing robustness to unseen data through empirical risk minimization.

Deep Neural Networks (DNNs) have revolutionized fields such as speech recognition and image classification by extracting complex patterns from data through their interconnected layers. However, DNNs face challenges including overfitting and uncertainty quantification, especially with limited datasets. Bayesian Neural Networks (BNNs) address these issues by employing probability distributions over weights, managing uncertainty and improving generalization in data-scarce scenarios. Techniques like Flipout enhance BNN training efficiency by decorrelating gradient estimates, improving convergence and uncertainty quantification.

In high-stakes applications, such as healthcare and autonomous driving, where both accuracy and uncertainty are critical, BNNs demonstrate significant promise. Research by Nudel et al. (2021) and Kasim et al. (2022) highlights the effectiveness of machine learning and deep learning models in predicting medical outcomes, illustrating how advanced techniques like BNNs and feature selection can substantially enhance prediction accuracy. Despite the "black box" nature of deep learning posing interpretability challenges,

integrating deep learning with feature selection methods improves both model interpretability and effectiveness, advocating for their application in real-world scenarios. The early AI systems were more interpretable, but the rise of Deep Learning (DL) models, characterized by complex architectures like DNNs, has resulted in a perception of these systems as "black boxes" (Castelvecchi, 2016). This complexity creates interpretability and accountability challenges, as failures in these systems can have significant consequences (Hu et al., 2021; Lakkaraju et al., 2020). Consequently, there is a growing emphasis on developing model-agnostic explanation methods to provide insights into these opaque models without needing access to their internal mechanics (Ribeiro et al., 2016).

Key concepts related to interpretability encompass understandability, comprehensibility, interpretability, and explainability. Understandability refers to a model's capacity to be comprehensible to humans, while interpretability focuses on the human's ability to grasp the model's reasoning. Explainability involves actions that elucidate the model's processes for users (Guidotti et al., 2018; Lipton, 2018). Transparency is further categorized into different models based on their support for user understanding (Lipton, 2018).

Enhancing interpretability in machine learning (ML) can improve decision-making by ensuring fairness, robustness against adversarial attacks, and validating causal relationships (Hall, 2018). Interpretability methods are categorized into global approaches, which provide an overview of model behaviour, and local approaches that focus on specific predictions (Bratko, 1997; Hall et al., 2017).

Prominent model-agnostic explanation methods include LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive Explanations). LIME generates

simpler surrogate models to explain predictions made by complex models, focusing on localized behaviour (Ribeiro et al., 2016; Samek et al., 2021). SHAP values, derived from game theory, quantify each predictor's contribution to a model's predictions, allowing for a nuanced understanding of individual predictor influences (Lundberg & Lee, 2017; Molnar, 2022). Together, these methods provide crucial insights that enhance the interpretability and accountability of AI systems, particularly in high-stakes applications.

Model calibration techniques are essential for adjusting predicted probabilities to better reflect actual outcomes. These methods improve the reliability of probabilistic predictions, which is crucial in clinical settings where accurate probability estimates directly inform decision-making. The most commonly used calibration methods include Platt Scaling (Platt, 1999), Isotonic Regression (Zadrozny & Elkan, 2002), and Temperature Scaling (Guo et al., 2017). In the context of mortality prediction, the goal is to ensure that the model's predicted probabilities of mortality closely align with the true outcomes, a task that is complicated by factors such as class imbalance, skewed data distributions, and the high complexity of machine learning models. Most traditional calibration methods struggle with highly complex, deep learning-based models, especially in the presence of imbalanced datasets common in healthcare applications. This gap in model calibration has driven recent research towards incorporating calibration techniques into Bayesian models and other ensemble methods for better alignment of predicted probabilities with true outcomes.

## **CHAPTER III: METHODOLOGY**

### **RESEARCH DESIGN AND METHODS**

#### **3.1 Overview of the Research Problem**

This study investigates the prediction of in-hospital mortality among patients who underwent general surgery by utilizing various clinical and demographic variables derived from established comorbidity indices, namely the Charlson and Elixhauser indices. Previous research by Simard et al. (2018) indicated that integrating these indices improves the prediction of 30-day mortality when compared to using either index independently. The analysis incorporates a diverse range of additional variables, including surgical history within the past year, smoking status, and critical laboratory parameters such as total leukocyte count, urea, creatinine levels, liver function tests (e.g., total, direct, and indirect bilirubin, alkaline phosphatase, SGOT, SGPT), as well as electrolyte levels (sodium and potassium). The dataset also encompasses the American Society of Anaesthesiologists (ASA) classification, types of surgical procedures (including open cholecystectomy, hernioplasty, and more), and postoperative complications like surgical site infections, pulmonary and cardiac complications, urinary tract infections, sepsis, reoperation, and readmission rates. However, challenges related to privacy concerns hindered the collection of comprehensive data.

The dataset in this study is relatively small, comprising approximately 931 samples, with a notable imbalance in the outcome variable (in-hospital mortality). To address this limitation, the study employs synthetic data generation techniques, particularly focusing on advanced generative models. The machine learning community has recently emphasized

the development of such models to autonomously capture intrinsic patterns in real-world data. Generative Adversarial Networks (GANs) have gained prominence for their effectiveness in generating realistic synthetic data while maintaining privacy (Goodfellow et al., 2014; Radford, Metz, and Chintala, 2015). In parallel, Variational Autoencoders (VAEs), introduced by Kingma et al. (2014), provide a powerful approach for data synthesis by approximating the underlying distribution of the data.

While VAEs and GANs have become popular in deep learning for generating synthetic data, GANs are often seen as superior for producing realistic images. However, this advantage has not been uniformly observed in the generation of tabular data (Singh and Ogunfunmi, 2022). VAEs consist of two main components: an encoder that compresses the input data into a lower-dimensional latent space defined by multivariate normal distributions, and a decoder that reconstructs the data by sampling from these distributions (Kingma and Welling, 2019).

This research focuses on constructing a tabular dataset from medical records collected over one year (2016-2017) from a teaching institution. Due to the small size of the dataset, the study employs the Variational Autoencoder (VAE) for synthetic data augmentation to enhance model development. The dataset is divided into training (75%) and validation (25%) subsets, with data normalization applied in batches to mitigate the impact of extreme values during model training. The deep learning model is designed using Keras, with dropout techniques implemented to reduce overfitting.

Model performance is evaluated using a variety of metrics, including specificity, sensitivity (recall), positive predictive value (precision), and F1 score. Additionally, the area under the receiver operating characteristic curve (AUROC) are employed to assess the model's predictive accuracy. The performance of the deep learning model is compared against traditional machine learning classifiers, including logistic regression, K-Neighbors

Classifier, SVM, random forest, gradient boosting, and XGBoost, to ascertain whether the deep learning model can match or exceed their performance in identifying mortality risk. Moreover, the study explores the use of ensemble models combining Bayesian and probabilistic approaches to enhance true positive identification and improve the F1 score. To further refine the reliability of predictions, the probabilities obtained from these models are calibrated using techniques such as temperature scaling, Platt scaling, and isotonic regression. This multifaceted approach aims to advance understanding of how machine learning, particularly deep learning and generative models, can be leveraged to predict critical health outcomes in surgical patients.

### **3.2 Operationalization of Theoretical Concern**

Data were collected from patients who had undergone general surgery, including variables from the Charlson and Elixhauser comorbidity indices to predict outcomes such as in-hospital mortality. Simard et al. (2018) demonstrated that combining these indices provides better prediction of 30-day mortality compared to either index used alone. In addition to these indices, other variables included surgery within one year, recent smoking history, and various laboratory parameters such as total leukocyte count, urea, creatinine, and liver function tests (e.g., total, direct, and indirect bilirubin, alkaline phosphatase, SGOT, SGPT), as well as electrolyte levels (sodium and potassium). The data also included ASA classification, surgical procedures (such as Open Cholecystectomy, Hernioplasty, Herniotomy, Lithotomy, Pyeloplasty, Appendicectomy, Omentoplasty, Small Bowel Resection, Laparoscopic ligation of adhesions, Modified Radical Mastectomy, Hysterectomy, Prostatectomy, Diagnostic Laparotomy, Nephrectomy, Gastrectomy, Esophagectomy, Gastric outlet Obstruction, Generalized peritonitis, and others), and

postoperative complications like surgical site infections (superficial, deep, and organ-space), pulmonary and cardiac complications, urinary tract infections, sepsis, reoperation, and readmission rates. Privacy concerns posed challenges in collecting comprehensive data.

Since the dataset was relatively small, consisting of only around 931 samples with a significant imbalance in the targeted outcome, generating synthetic data emerged as the most feasible solution. Recently, the machine learning community has focused on creating advanced generative models that can autonomously capture intrinsic patterns within and across real-world data records. Among these, Generative Adversarial Networks (GANs) have shown remarkable success in generating highly realistic synthetic data while preserving privacy (Goodfellow et al., 2014; Radford, Metz, and Chintala, 2015). In 2014, Kingma et al. introduced the Variational Autoencoder (VAE), followed by Goodfellow et al.'s GAN architecture just a year later. Both models employ deep neural networks to synthesize data by approximating its underlying distribution, though they approach the task differently—GANs estimate the data distribution implicitly, while VAEs do so explicitly. VAEs and GANs are popular tools for generating synthetic data, particularly in deep learning (Mi, Shen, and Zhang, 2018). While GANs tend to outperform VAEs in generating highly realistic synthetic images, this advantage has not consistently extended to tabular data generation (Singh and Ogunfunmi, 2022). A VAE consists of two primary components: an encoder and a decoder. The encoder reduces the input data into a lower-dimensional latent space characterized by multivariate normal distributions, typically Gaussian (Kingma and Welling, 2019). Each Gaussian distribution is defined by a mean and standard deviation, which determine the center and spread of the distribution, respectively (Kingma and Welling, 2014). The decoder then reconstructs the data by

sampling latent vectors from these distributions, ultimately generating the final output (Kingma and Welling, 2014).

This study focused on constructing a tabular dataset from medical records collected over one year (2016-2017) from a teaching institution. Given the small size of the dataset, the synthetic data generation technique of Variational Autoencoder (VAE) was employed to augment the data for model development. The dataset was split into training (75%) and validation (25%) sets for all models. Data were normalized in batches to prevent misleading results caused by extreme values during forward and backward propagation. Keras layer initializations were used to set the initial random weights, and dropout was incorporated to mitigate overfitting during the development of the deep learning model.

The deep learning model's performance was evaluated using metrics such as specificity, sensitivity (recall), positive predictive value (precision), and F1 score. Additionally, area under the receiver operating characteristic curve (AUROC) were used to assess the model's performance. These results were compared with those obtained from machine learning classification techniques, including logistic regression, K-Neighbors Classifier, SVM, random forest, gradient boosting, and XGBoost. The objective was to determine whether a deep learning model trained on a small dataset could perform as well as, or better than, traditional machine learning classifiers in identifying mortality.

In addition to VAE-based generative probabilistic modelling, separate and ensemble models were developed using Bayesian and probabilistic models, with a focus on improving true positive identification and achieving a high F1 score. Probabilities from these models were calibrated using temperature scaling, Platt scaling, and isotonic regression to enhance the reliability of the predictions.

### **3.2.1. Feature Importance**



To enable interpretability, there are a variety of different interpretation methods. This study will concentrate on model-agnostic explainability techniques of Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations, commonly known as SHAP, to derive features of importance in the prediction of mortality since these are the most popular interpretation techniques for humans to understand decisions made by ML models.

### **3.2.2. Additional statistics**

The results are presented as mean and standard deviation (SD) for continuous variables and as frequencies for categorical variables. Correlation analysis was conducted to determine any significant relationships between variables. Univariate analysis was performed using a logistic regression test to identify significant variables. Additionally, a two-sided independent Student's t-test (with a significance threshold of  $p < 0.05$ ) was used to evaluate whether there was a statistical difference between values in the preoperative and postoperative periods. A p-value of less than 0.05 was considered to indicate statistical significance.

### **3.3 Research Purpose and Questions**

The primary objective of this research is to develop predictive models for in-hospital mortality among patients who have undergone general surgery. This study aims to evaluate the effectiveness of various machine learning techniques, including deep learning and traditional classifiers, in predicting mortality outcomes based on a comprehensive dataset enriched through synthetic data generation techniques.

**RQ1** Which approach, Machine Learning or Deep Learning, is more effective for predicting mortality in a small dataset supplemented with synthetic data

**RQ2** Whether Generational autoencoder can be used using variational autoencoder to correct for imbalance in training data with superior results?

**RQ3.** Can the variables of importance identified by the explainability techniques, Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP), provide consistent results in explaining and interpreting the model for predicting postoperative mortality?

### **3.4 Research Design**

The research employs a quantitative approach, utilizing a retrospective cohort design. The dataset was divided into training (75%) and validation (25%) subsets to facilitate model development and evaluation.

**3.4.1 Population:** The study focuses on patients who underwent general surgery within a teaching institution from 2016 to 2017.

**3.4.2 Data Sources:** Data were collected from medical records, including a variety of clinical and laboratory variables:

- **Comorbidity Indices:** Charlson and Elixhauser indices to assess patient comorbidities.
- **Clinical Variables:** Recent smoking history, ASA classification, surgical procedures (e.g., Open Cholecystectomy, Hernioplasty, etc.), and postoperative complications (e.g., surgical site infections, pulmonary complications, urinary tract infections).

- **Laboratory Parameters:** Total leukocyte count, urea, creatinine, liver function tests (total, direct, and indirect bilirubin, alkaline phosphatase, SGOT, SGPT), and electrolyte levels (sodium and potassium).

### 3.4.3 Data Challenges

- **Sample Size:** The dataset consists of approximately 931 samples, which presents challenges due to its relatively small size and the significant imbalance in the targeted outcome (in-hospital mortality).
- **Privacy Concerns:** Comprehensive data collection faced challenges due to privacy issues, necessitating careful handling of sensitive patient information.

### 3.4.4 Limitations of DataSet Size and Balance

To address the limitations of the dataset size and balance, Variational Autoencoders (VAEs) were employed to generate synthetic data. This approach helps in augmenting the dataset while preserving the privacy of the original data.

### 3.4.5 Model Development Machine Learning Techniques

A range of classification models were utilized, including: **Traditional Models:** Logistic Regression, K-Neighbors Classifier, Support Vector Machine (SVM), Random Forest, Gradient Boosting, and XGBoost and **Deep Learning Model:** A deep learning architecture leveraging VAEs for synthetic data generation was developed.

### 3.4.6 Model Evaluation Metrics

The performance of the predictive models was evaluated using several metrics including **Sensitivity and Recall:** To assess the model's ability to correctly identify true positives (i.e., patients who experienced in-hospital mortality)

**Specificity:** To measure the model's accuracy in identifying true negatives.

**Positive Predictive Value (Precision):** To evaluate the proportion of correct identifications.

**F1 Score:** To provide a balance between precision and recall.

**Area Under the Area Under the Receiver Operating Characteristic Curve (AUROC):**

To assess the model's ability to distinguish between classes.

### **3.4.7. Calibration Techniques**

To enhance the reliability of the predictions from the models, calibration techniques were applied, including:

- **Temperature Scaling**
- **Platt Scaling**
- **Isotonic Regression**

## **3.5 Population and Study Sample**

The study was retrospective with a population of patients in a small data set who underwent General Surgery procedures between 2016 to 2017, the data for which period was made available by authorities

### **3.5.1 Sample Size and Selection of Sample**

To calculate the sample size with a presumptive mortality rate of 5% based on the hospital's past experience, the following formula was used for sample size estimation for proportions:

$$n=Z^2*p*(1-p)/E^2$$

This formula is commonly used for estimating sample sizes in studies involving proportions (Daniel, 1999). The parameters are defined as follows:

- **n** = required sample size
- **Z** = Z-value (standard normal deviate) corresponding to the desired confidence level (e.g., 1.96 for 95% confidence)
- **p** = estimated proportion (mortality rate), which is 0.05 (5% in this case)
- **E** = margin of error (precision of the estimate)

**Steps:**

1. **Confidence Level:** A 95% confidence level was selected, with a corresponding Z-value of 1.96.
2. **Margin of Error (E):** The margin of error represents the maximum acceptable difference between the true population parameter and the estimate. For this study, the estimate of mortality was considered to be within  $\pm 2\%$  of the true proportion, so an EEE of 0.02 was selected.
3. **Calculation:** The values were then applied to the formula.

The sample size assumes a confidence level of 95% ( $Z = 1.96$ ), an estimated mortality rate of 5% ( $p = 0.05$ ), and a margin of error of 2% ( $E = 0.02$ ) the sample size is calculated as: 457 as under

$$n=(1.96)^2 * 0.05 * (1-0.05)/(0.02)^2 =457.2$$

Thus, a sample size of approximately **457 participants** is needed to estimate the mortality rate with a 95% confidence level and a 2% margin of error.

### **3.6 Participant Selection**

since deep neural networks typically require large amounts of data, all general surgeries performed between 2016 and 2017, for which data were available in medical records, were included in the study. The study was retrospective, analyzing a small dataset of patients who underwent general surgery procedures between 2016 and 2017, for which data were available for research. **All surgical records** from the specified period were included to ensure the most comprehensive dataset possible.

### **3.7. Instrumentation**

An Excel sheet was used to collect data on all variables from anonymous patient records. Data were sourced from the Medical Record department of a teaching Institution in India

### **3.8 Data Collection Procedure**

Data were collected from Medical Records of patients who underwent surgery between 2016 to 2017 in a teaching Institution in India. Data consisted of preoperative factors including comorbidities in the patient's profile as enumerated by Charlson and Elixhauser including binary variables of Surgery within one year, smoking within one year, alcohol Abuse, HIV+ve status, deficiency Anemia, Rheumatoid Arthritis, Cardiac Pulmonale, Diabetes Mellitus, Chronic Hypertension, Hypothyroidism, Liver disease, Metastasis, Obesity, Renal failure, Tumour, Myocardial Infarction, Bronchial Asthma, Cerebrovascular accident, Chronic lung disease, Chronic Liver disease, Hemiplegia, Moderately Severe Liver disease, Preoperative Renal Failure, type of surgeries, as are commonly performed in general surgery setup including Lap Cholecystectomy, Open

Cholecystectomy, Hernioplasty, Herniotomy, Lithotomy, Pyeloplasty, Appendicectomy, Omentoplasty, Small Bowel Resection, Laparoscopic ligation of adhesions, Modified Radical Mastectomy, Hysterectomy, Prostatectomy, Diagnostic Laparotomy, Nephrectomy, Gastrectomy, Esophagectomy , Gastric outlet Obstruction, Generalised peritonitis, postoperative factors which influence surgical mortality including Deep Surgical Site Infection, Organ Space Surgical Site Infection, Abdominal wall Dehiscence, Pulmonary Complications, Cardiac Complication, Deranged Kidney Function Test, Urinary Tract Infection, Postoperative renal failure, reoperation, Readmission and Sepsis, Continuous variables will include , Laboratory parameters, Total Leucocyte Count, Urea, Creatinine, Total Serum Bilirubin ,Serum Bilirubin Direct ,Serum Bilirubin Indirect, Alkaline Phosphatase, SGOT,SGPT, Sodium and Potassium, in preoperative and postoperative period and duration of admission . The nominal variable included the American Society of Anaesthesiologists (ASA) Classification. Input variables were features as above that are used as input in the development of a model to predict the outcome (in-hospital mortality). The entire dataset Data is categorized into two classes based on whether the patient will experience mortality as an outcome of surgery during the hospital stay in the postoperative period.

### **3.9 Data Analysis**

The study aimed to investigate the impact of various risk factors on mortality using a Deep Neural Network model, with a focus on identifying the key contributing variables rather than providing a precise quantification of their individual effects. The objective of this

study was to develop a machine-learning prediction model for in-hospital mortality following non-cardiac surgery. Although missing values were initially considered for imputation, the proportion of records with missing data was negligible. As a result, these records were excluded from the analysis instead of applying imputation techniques.

### **3.9.1 Data Analysis Strategies**

The dataset for the study population, which included patient characteristics, comorbidities, preoperative and postoperative investigations, and complications, was organized into a CSV file. Machine learning models were built using Python 3.9 (Python Software Foundation, Wilmington, DE), utilizing the scikit-learn library and TensorFlow version 2.13.0. These models were trained to predict outcomes on the test set and evaluated based on metrics such as specificity, sensitivity (recall), positive predictive value (PPV or precision), and F1 score, which is the harmonic mean of precision and recall (calculated as  $2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$ ). Additionally, receiver operating characteristic (ROC) curves were generated, and the area under the curve (AUC) was computed for each model.

Probabilistic models, including a Variational Autoencoder (VAE) and Bayesian models, were developed both individually and in ensemble configurations, focusing on accurately identifying true positives. The predicted probabilities of these models were calibrated using methods such as temperature scaling, Platt scaling, and isotonic regression. Furthermore, model-agnostic explainability techniques, including Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), were employed to identify the most important features influencing mortality predictions.



### **3.10 Research Design Limitations**

#### Limitations of the Study

Limitations of the study include firstly the Sample Size, although the study utilizes synthetic data generation to augment a dataset of approximately 931 samples, the original sample size remains relatively small. This may limit the generalizability of the findings to broader patient populations and reduce the robustness of the predictive models. Secondly, imbalanced Data, where in dataset exhibits a significant imbalance between the number of positive (in-hospital mortality) and negative outcomes, this imbalance can lead to biased model performance, where classifiers may struggle to accurately identify the minority class, potentially affecting sensitivity and recall, thirdly Data Quality and Completeness which were collected from medical records, may be subject to inaccuracies, missing values, or inconsistencies. Such data quality issues can impact the reliability of the predictive models and the validity of the results, fourthly, privacy issues posed challenges in the comprehensive collection of data, which may have limited the inclusion of relevant clinical variables that could enhance model performance, fifthly Generalizability, the study was conducted within a single teaching institution, which may limit the applicability of the findings to other healthcare settings. Variations in patient demographics, surgical practices, and clinical protocols across institutions could affect outcomes, sixthly, Model Complexity, while advanced models like Variational Autoencoders (VAEs) and ensemble methods are employed, their complexity may lead to overfitting, especially given the small dataset size. Careful validation is necessary to ensure that models generalize well to unseen data, seventhly, Evaluation Metrics, although a variety of evaluation metrics are used,

relying solely on these metrics may not capture all aspects of model performance. For example, high precision does not necessarily imply high recall and trade-offs between different metrics can complicate the assessment of model efficacy. Eighthly, assumption of Data Distribution, while generative models like VAEs assume that the underlying data distribution can be approximated effectively, if the true distribution of the data is complex or differs significantly from the assumed model, the generated synthetic data may not adequately represent real-world scenarios, ninthly, Calibration Limitations, while calibration techniques are applied to improve prediction reliability, their effectiveness can vary based on the model and data characteristics. Inadequate calibration may still lead to suboptimal prediction probabilities, affecting clinical decision-making, tenthly, Temporal Considerations, since the data was collected from a specific time frame (2016-2017), and changes in surgical practices, patient demographics, or advancements in medical technology since then may affect the relevance of the findings to current clinical settings. These limitations highlight the need for a cautious interpretation of the study results and suggest avenues for future research to address these challenges and enhance predictive modeling in surgical outcomes.

## CHAPTER IV:

### RESULTS

A statistical analysis of quantitative data comprising of mean, standard deviation, minimum, maximum proportion, and standard error for Qualitative data is attached as Appendix “A “.A correlation analysis showed significant correlation of Alcohol Abuse with Smoking within one year of surgery(.45),Hypertension with Diabetes (.24),Obesity with Chronic Pulmonary disease(.12) and Hypertension(.17),Renal failure with deficiency Anaemia(.23),Bronchial Asthma with Chronic Pulmonary Disease(.26), Cerebrovascular Accident with Chronic Hypertension(.17),Chronic Liver disease with Alcohol abuse within one year(.12) , Pre OP Total Leukocyte Count (.19) and deficiency Anaemia(.23),Moderately severe Liver disease with Surgery within one year(.16) and Chronic Pulmonary Disease(.21),Preop Total Bilirubin with Surgery within one year(.13),Chronic Liver disease(.41),Moderately Severe Liver Disease (.31) and Pre Op Creatinine (.10), Preop SGOT with Preop Bilirubin Total(.20), Preop SGPT with Preop Bilirubin Total(.25), Preop SGOT(.78), Post op Leukocyte Count with Surgery within one year(.15),Post op Creatinine with Deficiency Anaemia(.14),Postop Urea(.56),Post OP SGOT with Surgery within one year(.10),Preop SGOT(.12),Postop Urea (.14),Postop Creatinine(.16), Postop Bilirubin Total(.16), Post OP SGPT with Surgery within one year(.21),Postop Creatinine (.11), Postop Bilirubin Total(.17), Post OP SGOT(.57), Postop Potassium with Preop Potassium(.10),Postop Creatinine(.11),Deep Surgical Infection with Deficiency Anaemia(.10), CVA(.15),Postop TLC(.12) Organ Space Infection with

Chronic Pulmonary Disease (.11),Wound Dehiscence with Moderately severe Liver Disease (.16),Gastric Outlet Syndrome with Surgery within one year(.12) and wound Dehiscence (.11),Generalised Peritonitis with Surgery within one year(.11),Pulmonary Complication with Deficiency Anaemia(.13),CVA(.19),Chronic Liver Disease (.11),Preop Urea (.12), Preop Creatinine (.13),Postop TLC(.14), Postop Urea(.10), Postop Creatinine(.10), Cardiac Complication with Smoking within one year of surgery(.11) , PreopCreatinine (.10), Chronic Pulmonary Disease(.10),Postop Creatinine (.12) and Pulmonary Complications (.23),Sepsis with Surgery within one year(.13), Chronic Pulmonary Disease(.11) , Deficiency Anaemia(.13), Chronic Liver Disease(.14), Pre Op Urea(.13 ) , PreopCreatinine(.17 ) , Pulmonary Complications(.40) and Deranged Electrolytes(.27). ,Duration of Stay with Surgery within one year(.10 ) , Smoking within one year of surgery( .14) , Alcohol Abuse(.14),Wound Dehiscence(.20),Gastric outlet Syndrome(.16),Generalised Peritonitis(.15), Pulmonary Complications(.30), Cardiac Complications(.15) and UTI(.31), Reoperation with Surgery with in one year(.12), Pre Op Urea(.11 ) , PreopCreatinine(.14 ) ,Preop Sodium(-.15),Preop SGOT(.12),Preop TLC(.12),Wound Dehiscence(.16),Gastric outlet Syndrome(.38),Generalised Peritonitis(.12),Pulmonary Complications(.10),Cardiac Complications(.17),Deranged Electrolytes(.15) UTI(.14),Sepsis(.14),Duration of stay(.24) , Readmission with Wound Dehiscence(.55),Reoperation(.12),Pre op Renal Failure with Deficiency Anaemia(.13),Preop Urea(.33),Preop Creatinine(.72),Preop Sodium(-.26),Postop SGOT(.28),Postop SGPT(.19), Post Op Urea(.40 ) and Postop Creatinine(.77 ) , Pulmonary Complications(.20), Cardiac Complications(.14),Electrolyte

Derangement(.69),Sepsis(.26),Reoperation(.15),Postop Renal Failure(.27), Post op Renal Failure with Deficiency Anaemia(.12), Pre op Creatinine(.17), Post Op Urea(.40 ),Postop Creatinine(.77 ), Pulmonary Complications(.16), Cardiac Complications(.22),Electrolyte Derangement(.72),Sepsis(.26), Death with Chronic Pulmonary Disease(.12) and Diabetes Mellitus(.10),CVA(.13),Chronic Liver Disease (.11), Preop Urea(.20),Preop Creatinine(.17),Preop Sodium(-.13). Postop TLC(.19), Post Op Urea(.25) and Postop Creatinine(.25), Pulmonary Complications(.41), Cardiac Complications(.15),Electrolyte Derangement(.28),UTI(.10),Sepsis(.54),Reoperation(.28),PostopRenal Failure(.23),Preop Renal Failure(.18),Lapcholecystectomy(-.16),Herniotomy(-.10),Omentoplasty(.14),Small Bowel resection(.43) Preop Urea with Preop TLC (.12),Pre op Creatinine with Chronic Pulmonary Disease(.20), Pre OPTLC(.17) and Preop Urea(.44),Pre op Bilirubin Total with Chronic Liver Disease(.41),Moderately Severe Liver Disease(.31), Preop TLC (.17) and Preop Creatinine(.10),Preop SGOT with Chronic Liver Disease(.14),Pre Op Urea(.17 ) and PreopCreatinine(.27 ), Preop SGPT with Chronic Liver Disease(.15),Moderately Severe Liver Disease (.13 ),Pre op TLC(.11) ,Pre op Creatinine (.20),Postop TLC with Chronic Liver Disease(.10),CVA (.10) and Preop TLC(.26),Post op Urea with Preop TLC(.10), Preop Urea(.11 ) and Preop Creatinine(.14),Postop Creatinine with Preop TLC(.15), Preop Creatinine(.10) , Preop Urea(.20) and Postop Urea, Postop Bilirubin Total with Chronic Liver Disease(.28),Deranged Electrolyte with Preop Urea(.26),Preop Creatinine(.52), Preop Sodium(-.23),Preop SGOT(.19), Preop SGPT(.12), Postop Urea(.35), Postop Creatinine(.58), UTI with Preop TLC(.10),Postop Creatinine(.10). A few examples of Important correlation in variables cluster to note are Hypertension and diabetes (0.24),

presence of both increases the risk of complications and mortality. Renal Failure and Deficiency Anemia (0.23) with potential for adverse outcomes. Bronchial Asthma and Chronic Pulmonary Disease (0.26) with the possibility of requiring respiratory support post-surgery, High correlation of Pre-op Renal Failure with Deficiency Anaemia(.13), Preop Urea(.33), Preop Creatinine(.72), Preop Sodium(-.26), Postop SGOT(.28), Postop SGPT(.19), Post Op Urea(.40 ) and Postop Creatinine(.77) suggesting that pre-surgery renal function is a strong predictor of post-surgery outcomes of Renal failure and so on. A two-sample t-test Analysis was done to find the statistical difference in quantitative variables. Differences between Preop TLC and Postop TLC, Preop Sodium and Postop Sodium levels were significant which suggests that surgery significantly alters these parameters, potentially impacting recovery and outcomes. Univariate analysis with death as the dependent variable in logistic regression showed Variables with statistical significance at a 5% level in Univariate Analysis were Distance from Hospital within 50 km, Deficiency Anaemia, Chronic Pulmonary disease, Diabetes Mellitus, Metastasis, Renal Failure, Chronic Liver disease, Preop TLC, Pre-op Urea, Preop Creatinine, Pre-op Sodium, Preop Direct Bilirubin, Omentoplasty, Small Bowel resection, Post OP TLC, Post op Urea, Post op creatinine, Post op Sodium, Post op Bilirubin Total, Postop bilirubin Direct, Post op bilirubin Indirect, Post op SGOT, Post op SGPT, Deep Surgical Site Infection, Organ Space Surgical site infection, Dehiscence, Pulmonary Complication, Cardiac complication, deranged KFT, UTI, Sepsis, Post op Renal Failure, Duration of stay and reoperation, There was a significant difference in TLC after Surgical intervention with a mean decrease of 824.37, and Sodium level with a mean decrease of .663. There was no

significant change in values of Preoperative and Postoperative Urea, Creatinine, Hepatic enzymes i.e. SGOT, SGPT, and Alkaline Phosphate, and None of the Bilirubin measurements. While overall there was no difference, the same cannot be generalized to individual data.

#### **4.1 Research Question One**

**Which method whether Machine Learning or Deep Learning is most capable of identifying mortality in a small dataset complemented with synthetic data?**

In this study, the small dataset consisting of 932 samples was split into training and evaluation sets in a 3:1 ratio. The training data exhibited class imbalance, with the target feature of mortality showing 658 instances of no mortality and 39 instances of mortality. The strategy used to balance data was oversampling by upsizing the minority class(Drummond and Holte, 2003). Variational Autoencoders were tried to correct the imbalance in training data with the generation of 619 samples to find which method yields better results. The parameters used to measure the performance of models with or without correction of class imbalance included accuracy, F1 score, recall, precision, true positive, and false negative. Table 1 and Table 2 show the results with and without correction of class imbalance.

**Table 4.1****Performance of Machine learning models without correction of class imbalance (Original Data)**

Data form	Model	accuracy	f1 score	recall	precision	true +ve	false -ve
Original	Logistic regression	0.88	0.47	0.92	0.32	12	1
Original	K neighbors classifier	0.96	0.38	0.23	1	3	10
Original	SVC	0.96	0.44	0.31	0.8	4	9
Original	Decision Tree classifier	0.91	0.40	0.53	0.31	7	6
Original	Random forest classifier	0.94	0.14	0.07	1	1	12
Original	Gradient boosting classifier	0.96	0.67	0.62	0.73	8	5
Original	XGB classifier	0.95	0.59	0.62	0.57	8	5

Looking at the comparison (Table 2), the Deep Learning model trained with VAE-augmented data emerged as the best performer, offering the highest F1 score and a balanced recall and precision. The application of VAE data augmentation generally improved the performance of most models, particularly for complex models like Decision Trees, Random Forests, Gradient Boosting, and XGBoost. However, simpler models like Logistic Regression and SVC struggled with the VAE-augmented data.



**Table 4.2 Comparison of Machine learning model with Deep Learning Model with data imbalance corrected from Data Augmentation by Variational Autoencoder (VAE)**

Data Form	Model	Accuracy	F1 Score	Recall	Precision	True +ve	False -ve
Variational AE	LogisticRegression	0.86	0.27	0.46	0.19	6	7
Variational AE	KNeighborsClassifier	0.93	0.47	0.53	0.41	7	6
Variational AE	SVC	0.93	0.12	0.07	0.33	1	12
Variational AE	DecisionTreeClassifier	0.91	0.47	0.69	0.36	9	4
Variational AE	RandomForestClassifier	0.96	0.66	0.53	0.87	7	6
Variational AE	GradientBoostingClassifier	0.96	0.64	0.61	0.67	8	5
Variational AE	XGBClassifier	0.96	0.71	0.76	0.66	10	3
Variational AE	Deep Learning Model	0.96	0.71	0.85	0.61	11	2

The results indicate that while the DecisionTreeClassifier, RandomForestClassifier, GradientBoostingClassifier, and XGBClassifier show improved performance with variational Autoencoder transformed data, **Logistic Regression** and **SVC** do not perform well on the VAE data, indicating that simpler models may not be able to capture the complex patterns that the VAE is generating. The **Deep Learning model**, **Random Forest**, and **XGBoost** models can effectively handle the transformed feature space created by the VAE. Several models, especially Random Forest and KNN on the original data, struggled

with the imbalanced nature of the dataset. The VAE transformation appears to help mitigate some of these issues, particularly improving recall for these models. The Generative, Deep Neural Network (DNN) outperforms them all in terms of identifying True positives, achieving the best performance.

## **4.2 Research Question Two**

**Whether Generational autoencoder can be used using variational autoencoder to correct for imbalance in training data with superior results?**

The dataset utilized in this study is both limited in size and significantly skewed, given that mortality events are infrequent. In binary classification scenarios where the target class is highly imbalanced, standard machine learning algorithms tend to focus on maximizing overall accuracy. As a result, these algorithms often classify the majority of instances as belonging to the more prevalent class. This scarcity of examples for the minority class can hinder the learning algorithm's ability to effectively capture the characteristics of this class, leading to suboptimal performance in predictive accuracy (Japkowicz & Stephen, 2002). Consequently, this results in low recall for the minority class, which is usually the class of primary interest. To address the issue of imbalanced data, various techniques are available. This study explores how different oversampling methods can affect the performance of a Deep Learning model on an imbalanced dataset, specifically investigating whether the Variational Autoencoder (VAE) can yield superior outcomes. Two oversampling techniques were implemented: SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002) and its derivatives, alongside the VAE (Kingma et al., 2014). Similar

to SMOTE, the VAE generates new samples that are akin to those in the original dataset, albeit with slight variations (Zhang et al., 2018). While Random oversampling resulted in an F1 score of .61, the variational autoencoders method could achieve an F1 score of .77. Results are shown in Table. The table provides the performance metrics of a Deep Learning model under different data augmentation techniques, specifically various oversampling methods. Each row represents the results obtained by applying a different oversampling technique to address class imbalance in the dataset. The metrics include Accuracy, F1 Score, Recall, Precision, True Positives (TP), and False Negatives (FN) and ROC-AUC Score. The table below shows the results:

**Table 4.3. Data augmentation techniques and performance of DNN models**

Data Form Oversampling	Model	Accuracy	F1 Score	Precision	Recall	True +ve	False -ve	ROC-AUC Score
Random oversampling	Deep Learning	.96	.64	.67	.62	8	5	.93
SMOTE	Deep Learning	.95	.65	.56	.77	10	3	.86
BSMOTE	Deep Learning	.97	.71	.67	.77	10	3	.92
Adasyn	Deep Learning	.94	.59	.48	.77	10	3	.91
Deep Smote	Deep Learning	.96	.75	.63	.92	12	1	.97
Variational Autoencoder	Deep Learning	.96	.71	.61	.85	11	2	.95

Random oversampling leads to high precision but lower recall, but a relatively low F1 score (0.64). SMOTE improves recall (.77) but precision drops to .56 with an F1 score of 0.65. BSMOTE(Borderline-SMOTE) shows an F1 score of .71, and ROC-AUC is also higher (.92) showing better distinguishing between classes. Adasyn results in recall (0.77), meaning it misses many positive cases, and precision drops (0.48), which indicates a higher number of false positives. Deep SMOTE achieves a strong performance across the board, with high recall (0.92) and precision (0.63). The ROC-AUC score of 0.97 is the highest, indicating excellent overall model performance in separating the classes. The VAE-based model showed consistently high performance across multiple metrics. VAE transformed data produced a high F1 score of 0.71, a strong recall of 0.85, and an ROC-AUC score of 0.95. It also identifies 11 true positives and only 2 false negatives, indicating robust detection of the minority class. This shows that the model based on VAE-transformed data is effective at identifying the minority class while maintaining high accuracy and precision. The ROC-AUC of 0.95, though slightly lower than Deep SMOTE, is still very good. While both the Deep SMOTE and Variational Autoencoder models achieve comparable performance in terms of F1 score and recall, the VAE offers several additional advantages that make it a superior choice for mortality prediction. First, the VAE's ability to model complex latent representations allows it to uncover hidden patterns in the data, making it more adaptable to diverse datasets and robust to noise. Secondly, the generative capabilities of the VAE provide a novel approach to handling class imbalance and uncertainty, offering more flexibility for research and real-world applications. Thirdly, while Deep SMOTE

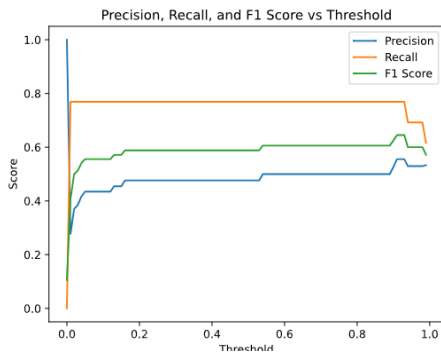
directly augments the dataset to handle imbalance, VAEs offer a different approach by learning a latent representation of the data distribution, allowing for complex generative capabilities (Kingma & Welling, 2014). VAE's ability to generate synthetic samples that capture the underlying distribution of the minority class is particularly useful in medical datasets where the occurrence of death may be rare compared to survival. Given the critical nature of death prognostication, where missing a true positive can have severe consequences, the VAE model is the most appropriate choice. Lastly, VAEs are at the forefront of modern deep learning research and have been widely used for generative tasks, anomaly detection, and representation learning. VAE-transformed data allows leveraging state-of-the-art machine learning techniques to tackle a challenging imbalanced classification problem.

DNN model with Autoencoder for generation of synthetic data gave a probabilistic generative output which varied with each run creating high uncertainty in predicting the outcome of mortality with no definite bounds of prediction at individual level. At a broader level prognostic in health management face uncertainty problem which comes from the effects of two critical uncertainties: 1) epistemic uncertainty, accounting for the uncertainty in the model, and 2) aleatoric uncertainty, representing the impact of random disturbance, such as measurement errors (Kendall, Gal, 2017). To measure the effects of uncertainty, stochasticity is a key component of many modern neural net architectures and training algorithms. The most widely used regularization methods are based on randomly perturbing a network's computations. Wen et al. (2018) described Flipout, a technique for probabilistic modeling that reduces gradient variance during training. Regularization methods often involve perturbing weights, as in Bayesian neural networks, or applying

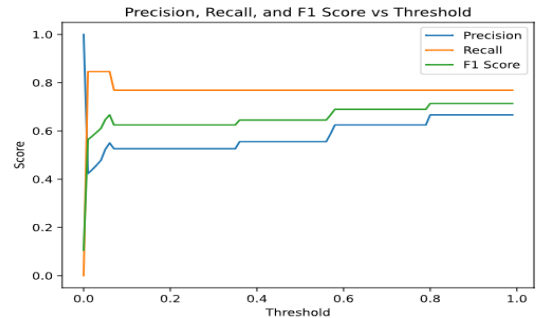
stochastic techniques like dropout to activations. Flipout specifically introduces independent weight perturbations to decorrelate gradients, but perturbing weights often incurs additional computational costs. There are (typically) many more weights than there are units in a neural network, so computing the weight perturbation for every single element in a mini-batch becomes incredibly costly on computational resources, As a result, methods that are regularized by weights will usually only use a single sample per mini-batch (Wen, et al.,2018).

For Model development for prognostication of mortality, two approaches were tried. In the first approach, individual models were trained, which included a DNN model with VAE, a **probabilistic generative model**, two Probabilistic models with a Flipout layer at the end versus in all layers and a Bayesian model. To generate the bounds the models were run 200 times and results were aggregated.

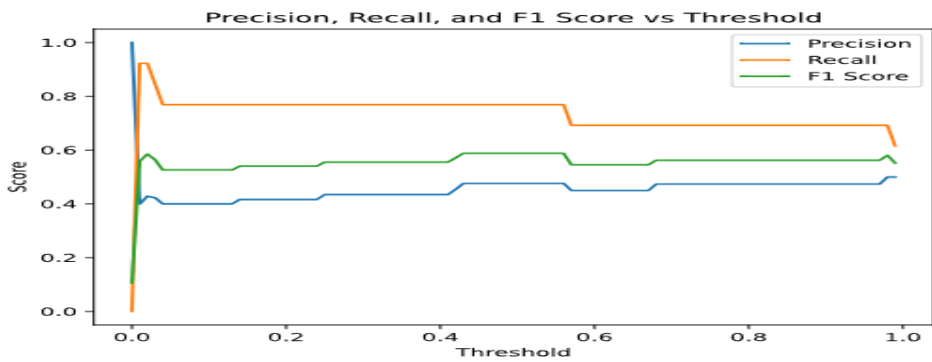
Smote



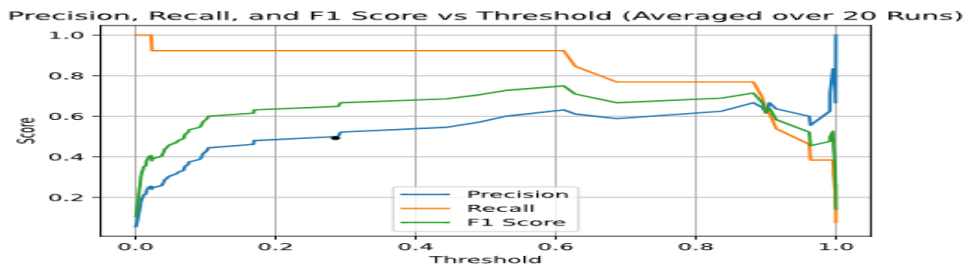
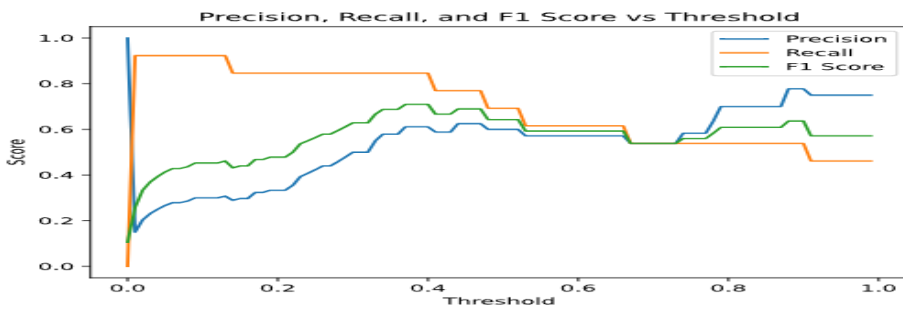
BS Smote



Adasyn



VAE



**Fig 4.1 Comparison of SMOTE and Its Variants**

This comparison illustrates the effectiveness of various oversampling techniques on imbalanced datasets. VAE demonstrates superior performance in handling data imbalance,

while SMOTE and its variants show varying levels of improvement based on precision, recall, and F1 score.

DNN model with Autoencoder for generation of synthetic data gave a probabilistic generative output which varied with each run creating high uncertainty in predicting the outcome of mortality with no definite bounds of prediction at individual level. At a broader level prognostic in health management face uncertainty problem that comes from the effects of two critical uncertainties: 1) epistemic uncertainty, accounting for the uncertainty in the model, and 2) aleatoric uncertainty, representing the impact of random disturbance, such as measurement errors (Kendall, Gal, 2017). To assess the impact of uncertainty, stochasticity plays a crucial role in many contemporary neural network designs and training methods. Regularization techniques often incorporate stochasticity by introducing random perturbations to the network's computations, which helps prevent overfitting and promotes better generalization. One such method, introduced by Wen et al. (2018), is Flipout, a regularization approach specifically designed for probabilistic modelling.

Flipout operates by introducing decorrelated perturbations to a network's weights, unlike traditional Bayesian neural networks, which perturb weights in a correlated manner. The challenge with weight perturbation is that neural networks typically have far more weights than units, making it resource-intensive and computationally expensive to calculate weight perturbations for every element in a mini-batch. This limitation often forces weight-regularized methods to utilize only one sample per mini-batch, hindering their ability to learn complex patterns (Wen et al., 2018).



In contrast, Flipout enables more efficient computations by allowing the network to leverage mini-batches effectively without the high computational costs associated with weight perturbation. By introducing stochasticity through decorrelated weight perturbations, Flipout enhances the model's ability to explore a wider range of parameter variations, thus improving generalization. This efficiency not only enhances the scalability of probabilistic models but also promotes better generalization by incorporating a broader variety of samples during training (Wen et al., 2018). Therefore, Flipout presents a promising solution to address some of the inherent limitations faced by conventional regularization techniques in probabilistic neural networks.

For Model development for prognostication of mortality, two approaches were tried. In the first approach, individual models were trained, which included a DNN model with VAE, a probabilistic generative model, two Probabilistic models with Flipout layer at the end versus in all layers, and a Bayesian model. To generate the bounds the models were run 200 times and results were aggregated.

**Table 4.4 Comparison of Probabilistic Models**

Probabilistic Model	Accuracy	Precision	Recall	F1 score	True -ve	False -ve	False +ve	True+ve	ROC-AUC Score	Mean P of where class is 1 & SD and mean p of positive class & SD	Global Entropy and /Entropy of positive class
VAE	.97	.77	.77	.77	217	3	3	10	.91 at threshold of .47	.64+/- .38 .06+/- .20	.03 / .10

Flipout Last layer	.96	.67	.62	.64	216	4	5	8	.84	0.58+/ -.32 .21+/- .15	5.3/ 2.3
Flipout All layers	.06	.06	1	.11	0	220	0	13	.30	.52 +/-0.009 52+/-0.001	161.16/8. 99
Bayesian	.96	.70	.54	.61	217	3	6	7	.80	.63+/-0.009 .55+/-0.003	109.44/5. 27

### VAE Model:

This model had strong performance across all metrics, with a balanced precision, recall, and F1 score of 0.77. The model achieved high accuracy (0.97), indicating that it correctly predicted most of the cases. The ROC-AUC score of 0.95 suggests excellent discriminatory ability between the classes. With a **Global Entropy (0.03)**, the model was generally confident in its predictions. **Positive Class Entropy (0.10)** means there is more uncertainty in the predictions for the positive class compared to the overall dataset, reflecting the model's challenge in predicting the minority class with high certainty. **Mean Probability for Class 1 at 0.64 means that** in predicting the positive class (Class 1), the average probability the model assigns to this class is approximately 0.64 which indicates that when the model identifies a sample as belonging to Class 1, it does so with fairly high confidence, as the probability is well above 0.5. The model assigns a **low overall probability to the**

**positive class(.06)**, which means that in most cases, it predicts the negative class (class 0). However, for samples where the true class is 1, the model is much more confident. At the threshold of 0.47, the model's global entropy is very low (0.03 for global entropy, 0.10 for positive class entropy). This indicates that the model's predictions are generally confident, with minimal uncertainty, especially when it predicts positive cases. A low entropy means the model is not "uncertain" about its classification and is making more deterministic decisions.

#### **Flipout (Last Layer):**

The model underperforms compared to the VAE, with a significant drop in precision (.67), recall (.62), and F1 score (0.64). The accuracy is still relatively high (0.95), but the lower precision and recall suggest that it may struggle with classifying the positive class correctly. With **ROC-AUC of 0.84**, it is less capable of distinguishing between the classes compared to VAE. The **mean probability of the positive class** is relatively low ( $0.21 \pm 0.15$ ), indicating some uncertainty or low confidence in predicting the positive class, which could explain the lower recall. **Global entropy** values show that there is some uncertainty in predictions, though it is better compared to the **Flipout All Layers** model.

#### **Flipout (All Layers)**

The **Flipout All Layers model** performs poorly with **extremely low accuracy (0.06)** and **precision (0.06)**, although its **recall** is 1.0. This means that while it identifies all the true positives, it does so at the expense of many false positives, resulting in poor overall performance. The **F1 score** of 0.11 reflects this imbalance, with very low precision but

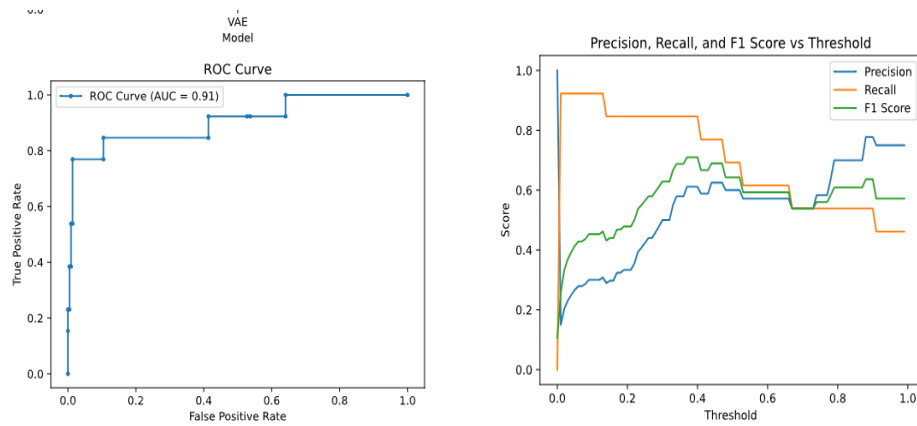
perfect recall. The **ROC-AUC** of 0.30 suggests that the model does not differentiate between the two classes well, essentially performing no better than random guessing. The **mean probability for both the positive class and Class 1** is very stable ( $0.52 \pm 0.0009$ ), which suggests the model is assigning similar probabilities across instances, possibly due to overfitting or lack of sufficient uncertainty propagation through the layers. **High global entropy values (161.16)** indicate significant uncertainty across the model, which suggests that adding Flipout to all layers introduces too much stochasticity, leading to a breakdown in prediction accuracy.

#### **Bayesian Model:**

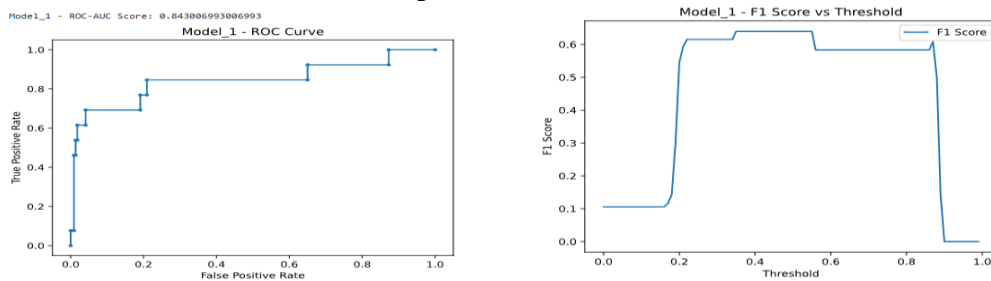
The Bayesian model performs similarly to the Flipout Last Layer model, with an accuracy of 0.96 and higher precision (0.70), but lower recall (0.54). This indicates that the model is better at avoiding false positives than it is at capturing all true positives, which was desirable considering that the objective was to predict mortality in clinical settings. The F1 score of 0.61 reflects this trade-off between precision and recall, showing slightly lower performance compared to the Flipout Last Layer model. ROC-AUC of 0.80 is slightly lower than that of the Flipout Last Layer model (0.84), suggesting a moderate ability to differentiate between classes. Mean probabilities for both Class 1 ( $0.63 \pm 0.009$ ) and the positive class ( $0.55 \pm 0.003$ ) are higher than in the Flipout Last Layer model, indicating more confidence in its positive predictions. Global entropy (109.44) is lower than the Flipout All Layers model but higher than the Flipout Last Layer model. This suggests that the Bayesian model incorporates uncertainty effectively without introducing excessive randomness.

Thus, the VAE model appears to be the most balanced, while the Flipout model and the Bayesian model both struggle with significant uncertainty. These two models either need further refinement or are not suitable for the prediction of mortality in the current context. VAE model, with its balanced precision and recall, high accuracy, and low uncertainty is a strong candidate for Deployment in hospitals for the prediction of mortality

### VAE Model

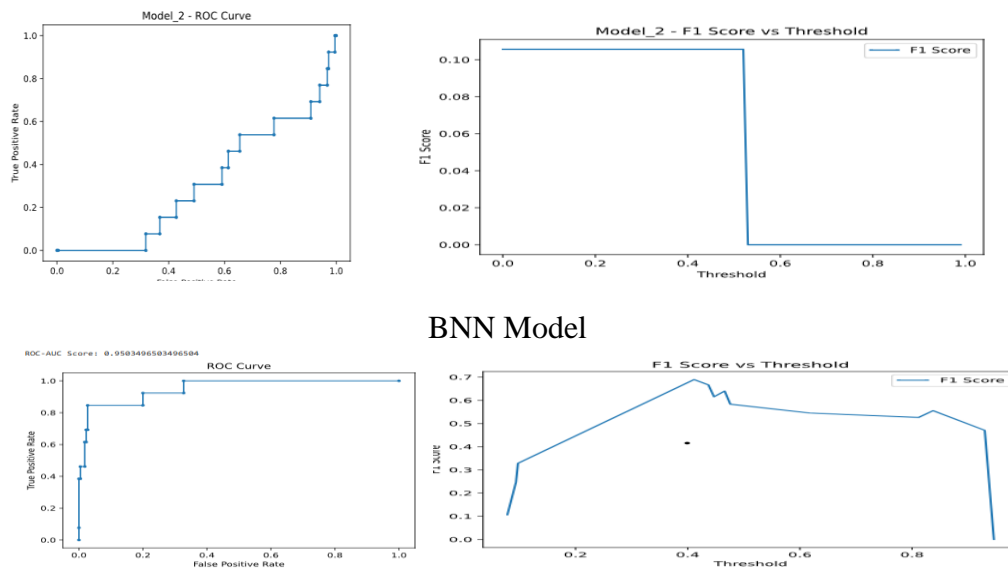


### Flipout at the End Model



### Flipout in All layers Model

Model\_2 – ROC-AUC Score: 0.30



**Figure 4.2 Comparison of ROC of VAE and Bayesian Models**

**Ensemble Modeling as a Strategy:**

Since probabilistic models did not perform well, and the aim was to move from a deterministic approach to a stochastic approach to better capture variability in data, explore a wider solution space, and provide probabilistic output, in the second approach Ensembles models were used where in a combination of VAE, Model\_1, Model\_2, Bayesian were combined and evaluated for accuracy, precision, recall, F1 score, True positive (TP), False positive (FP), False negative(FN), True positive (TP), Area Under the Curve, Mean probability where the outcome is mortality i.e. 1, mean global probability, global entropy and mean entropy where the outcome is 1(Table-4.5). The table uses the mean probability of class which is equivalent to the mean probability of class 1 and mean probability where the class is 1 which is more of an evaluation metric rather than a fundamental model output and is important in the context of **performance evaluation**, such as when model predicts

the positive class (class 1) and where ground Truth is also positive class. The details of the results are as under:

**Table 4.5 Comparison of Ensemble models**

Ensemble Configuration	Acc	Precision	Rec	F1	TN	FP	FN	TP	ROC AUC	Mean P of where the class is 1 & SD and mean p of positive class & SD	Enter (Global/Positive class
<i>VAE + Flipout at Last Layer+Bayesian_</i>	.97	.65	.85	.73	214	6	2	11	.86	.54+/-03 64'+-.06	5.4/2.5
<i>VAE+Bayesian+Flipout All layers</i>	.97	.65	.85	.73	214	6	2	11	.83	.63+/- 0.04 0.56 +/-02	5.4/2.5

VAE+Bayesian+Flipout All layers+Last layer	.97	.65	.85	.73	214	6	2	11	.90	.63+/-0.05 .56+/-0.05	5.4/2.5
VAE+Flipout Last Layer+ All layers	.97	.65	.85	.73	214	6	2	11	.94	.63+/-0.04 56+/-0.02.	5.4/2.55
Baysian +Flipout last layer+Flipout All layers	.96	.67	.62	.64	216	4	5	8	.90	.63+/-0.07 55+/-0.03	5.44/2.55
Baysian +Flipout last layer	.96	.67	.62	.54	216	4	5	8	.88	.63+/-0.03 .59+/-0.04	5.4/2.56
Baysian +Flipout All layers	.96	.67	.62	.64	216	4	5	8	.82	.55+/-0.16 .37+/-0.07	5.43/2.51
Flipout All layer+Flipout last layer	.96	.67	.62	.64	216	4	5	8	.82	.55+/-0.16 .37+/-0.07	5.43/2.38
VAE+Flipout_last_Layer	.96	.62	.77	.69	214	6	3	10	.95	.61+/-0.31 .13+/-0.16	76.81/5.3
VAE+Flipout_all Layer	.97	.77	.77	.77	217	3	3	10	.90	.57+/-0.02 .64+/-0.04	5.4/2.5
VAE+Bayesian	.97	.65	.85	.73	214	6	2	11	.89	.67+/-0.20 .13+/-0.16	5.4/2.8

### Pre-Calibration Summary of models

Of all the models, VAE +Flipout Last Layer had the ROC-AUC score (.95) and strong precision, recall, and F1 score, making it an excellent choice before calibration. **The**



**ensemble** had high recall with slightly lower precision, which is good for identifying more positive cases. Other models which had high performance were **VAE + Flipout All Layers** with a balance between high F1 score and a value of .77 across other matrices of precision and recall and ROC-AUC score at .90, **VAE+ Bayesian+ Flipout All layers+ Last layer** with equivalent matrices and ROC-AUC score of .90, **VAE+Flipout Last Layer+ All layers** with precision (.65 ).recall (.85), F1 score(.72) and ROC-AUC (.94 ), **Baysian +Flipout last layer+Flipout All layers** with precision of .67, recall of .62, F1 score of .64, and ROC-AUC value of .90, **VAE+Bayesian+ Flipout all** with precision of (.65), recall of .85), F1 score of (.73, ROC-AUC score of .90.

Identifying models that accurately predict the true underlying probabilities for each test case would be ideal. However, the challenge lies in our inability to train models to effectively estimate these probabilities. This difficulty can stem from a lack of knowledge about the appropriate parametric model type, the limited size of the training sample hindering accurate parameter estimation, or the presence of noise within the data. Typically, a combination of these issues manifests to varying degrees. Furthermore, we often lack access to the true underlying probabilities; we only know whether a case is positive or negative, complicating the task of determining if a model accurately predicts these probabilities (Caruana, 2004).

Guo et al. (2017) highlighted that the extensive capacity of neural network models, combined with their propensity to overfit complex datasets, makes them particularly susceptible to calibration problems. In many cases, standard deep learning approaches generate probabilistic predictions that are overly confident, especially when training sets

are small. This tendency exacerbates existing issues and can lead to undesirable outcomes when deep neural networks are implemented in situations requiring precise uncertainty quantification. Consequently, various calibration techniques, such as temperature scaling, Platt scaling, and isotonic regression, have been employed for ensembles.

Achieving well-calibrated probabilistic predictions is essential for effective risk management, especially when decisions hinge on the reliability of probabilistic model outputs. Calibration refers to a model's ability to generate probabilistic predictions that accurately reflect the true likelihood of various outcomes. Specifically, a model should provide a calibrated measure of confidence alongside its predictions; in other words, the probability linked to the predicted class label should mirror its actual likelihood of correctness. In models with good calibration, the predicted probabilities correspond with the observed frequency of events. For instance, if a model forecasts an event with a probability of 0.9, that event should occur 90% of the time over an extended period.

Rahaman (2021) emphasized that creating well-calibrated models is vital for fostering public trust in machine learning applications, particularly in AI-driven medical diagnostics, as it directly relates to the acceptance of new technologies. Common methods for calibrating deep learning models in low-data situations include ensembling, Platt scaling, temperature scaling, and mixup data augmentation (Rahaman, 2021).

. Accordingly, Ensembles of all the basic models were calibrated with temperature scaling, Isotonic regression, and Platt's scaling. However, calibration techniques have problems for two reasons firstly, most calibration methods assume a more balanced distribution of classes, and secondly, the model could be biased toward predicting the majority class

(zeros), which would result in poor calibration of the probabilities for the minority class (ones). Hence, calibrated probabilities were assessed using Brier's score to see how close are predicted probabilities with the actual before and after the calibration of models.

**Table 4.6 Evaluation of models as per matrices with Temp Scaling, Platt Scaling, and Isotonic regression**

Ensemble Configuration (calibrated with temp scaling)	Method of calibration	Accuracy	Precision	Recall	F1	TN	FP	FN	TP	Best Threshold	ROC AUC score
VAE+Flipout_last_Layer	Temp Scaling	.96	.62	.77	.69	214	6	3	10	.99	
	Platt Scaling	.94	.52	.76	.62	211	9	3	10	.08	.85
	Isotonic regression	.97	.67	.77	.71	215	5	3	10	.20	.87
VAE+Flipout_all_Layer	Temp Scaling	.97	.71	.77	.74	216	4	3	10	.58	
	Platt Scaling	.97	.77	.77	.77	217	3	3	10	.26	.87
	Isotonic regression	.97	.77	.77	.77	217	3	3	10	.24	.87

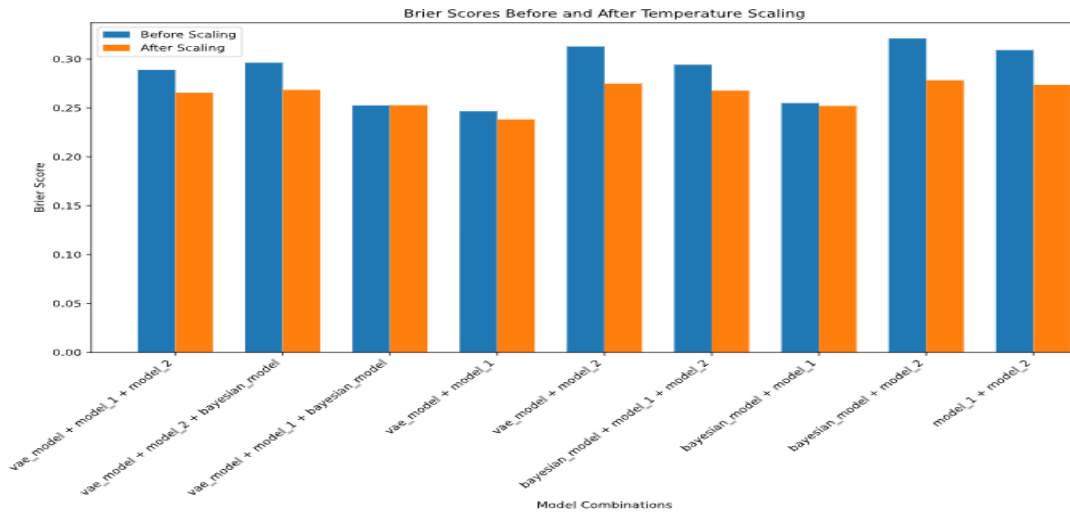
VAE + Flipout all Layer+Bayesian_	Temp scaling	.97	.77	0.77	0.77	217	3	3	10	.57	
	Isotonic regression	.97	.77	0.77	0.77	217	3	3	10	.50	.87
	Platt Scaling:	.97	.77	0.77	0.77	217	3	3	10	.21	.85
VAE + Flipout at Last Layer+Bayesian_	Temp scaling	.97	.77	.77	.77	217	3	3	10	.79	
	Isotonic regression	.97	.77	.77	.77	217	3	3	10	.25	.87
	Platt Scaling:	.97	.66	.77	.71	215	5	3	10	.14	.85
VAE+Flipout at Last+Flipout All layers	Temp scaling	.97	.77	.77	.77	217	3	3	10	.59	
	Isotonic regression	.97	.77	.77	.77	217	3	3	10	.50	.87
	Platt Scaling:	.97	.77	.77	.77	217	3	3	10	.20	.86
VAE+Flipout Last +Flipout All layers+Bayesian	Temp scaling	.97	.75	.69	.72	217	3	4	9	.58	
	Isotonic regression	.97	.69	.69	.69	216	4	4	9	.57	.83
	Platt Scaling:	.96	.62	.76	.69	214	6	3	10	.15	.84
VAE+ Bayesian_	Temp scaling	.97	.77	.77	.77	217	3	3	10	.78	
	Isotonic regression	.97	.77	.77	.77	217	3	3	10	.2	.87
	Platt Scaling:	.97	.77	.77	.77	217	3	3	10	.26	.90
Baysian +Flipout last layer+Flipout All layers	Temp scaling	.94	1	0	0	220	0	13	0	.5	.38

	Isotonic regression	.88	.10	.15	.12	202	18	11	2	.51	.40
	Platt Scaling:	.39	.07	.84	.13	80	140	2	11	.05	.55
Bayesian+Flipout Last Layer	Temp scaling	.06	.06	1	.11	0	220	0	13	.56	
	Isotonic regression	.94	.50	.15	.24	218	2	11	2	.6	.37
	Platt Scaling:	.69	.10	.61	.18	154	66	5	8	.06	.63
Bayesian+Flipout All Layer	Temp scaling	.06	.06	1	.11	0	220	0	13	.10	
	Isotonic regression	.08	.06	1	.11	5	215	0	13	.34	.51
	Platt Scaling:	.48	.07	.69	.13	104	116	4	9	.06	.52
Flipout last Last Layer+Flipout All Layer	Temp scaling	.06	.06	1	.11	0	220	0	13	0	
	Isotonic regression	.06	.06	1	.11	1	219	0	13	.04	.38
	Platt Scaling:	.94	.25	.08	.12	217	3	12	1	.39	

Temperature scaling consistently performed well across most ensemble models. It maintained high accuracy and a strong balance between precision and recall, making it a reliable choice for calibrating models. The results are shown in Table 4.7.

**Table 4.7 Calibration of ensemble models with Temperature Scaling and brier score before and after**

Model Combination	Brier Score (Before Scaling)	Brier Score (After Scaling)	Difference After Scaling	Observation
vae_model + model_1 + model_2	0.2889	0.2656	-0.0233	Moderate reduction in Brier Score, indicating improved probability calibration after scaling.
vae_model + model_2 + bayesian_model	0.2963	0.2685	-0.0278	Significant improvement after scaling. The initial Brier Score was high, indicating poorer performance.
vae_model + model_1 + bayesian_model	0.2526	0.2528	0.0002	Minimal change after scaling, indicating good probabilistic performance even before scaling.
vae_model + model_1	0.2466	0.2383	-0.0083	Lowest initial Brier Score, showing that this combination has good probability estimation pre- and post-scaling.
vae_model + model_2	0.3128	0.2749	-0.0379	The largest reduction in Brier Score after scaling, indicating substantial improvement in probability estimation.
bayesian_model + model_1 + model_2	0.2942	0.2678	-0.0264	Noticeable reduction, showing improved calibration after scaling.
bayesian_model + model_1	0.255	0.2521	-0.0029	Small improvement after scaling, indicating good initial calibration.
bayesian_model + model_2	0.3211	0.2783	-0.0428	Highest initial Brier Score, indicating poor calibration before scaling. Scaling substantially improved it.
model_1 + model_2	0.3092	0.2737	-0.0355	Large reduction in Brier Score after scaling, indicating that scaling helps to align predicted probabilities.



**Figure 4.3 Brier Score for Different Ensemble Models before and after Temperature Scaling**

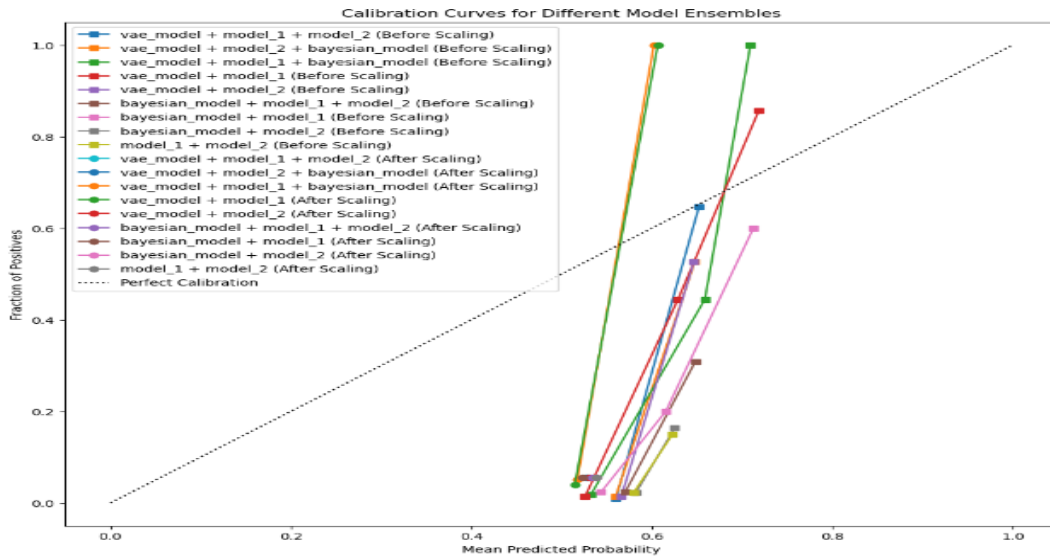
**Platt Scaling** struggled with certain models particularly those which involves probabilistic models involving **Bayesian + Flipout layers**) with poor performance in terms of precision and F1 score while it worked fine with simpler VAE-based ensembles. **Isotonic Regression** came out as relatively stable and provided competitive performance close to Temperature Scaling but tended to be slightly more conservative in probability estimation, leading to slightly lower precision in some cases. For most configurations, the best threshold post-calibration was generally lower with temperature scaling than with Platt or Isotonic methods. The AUC generally remained stable, with most calibrated models achieving scores between 0.85 and 0.90.

**Table 4.8 Brier Score (Before and After Isotonic Calibration)**

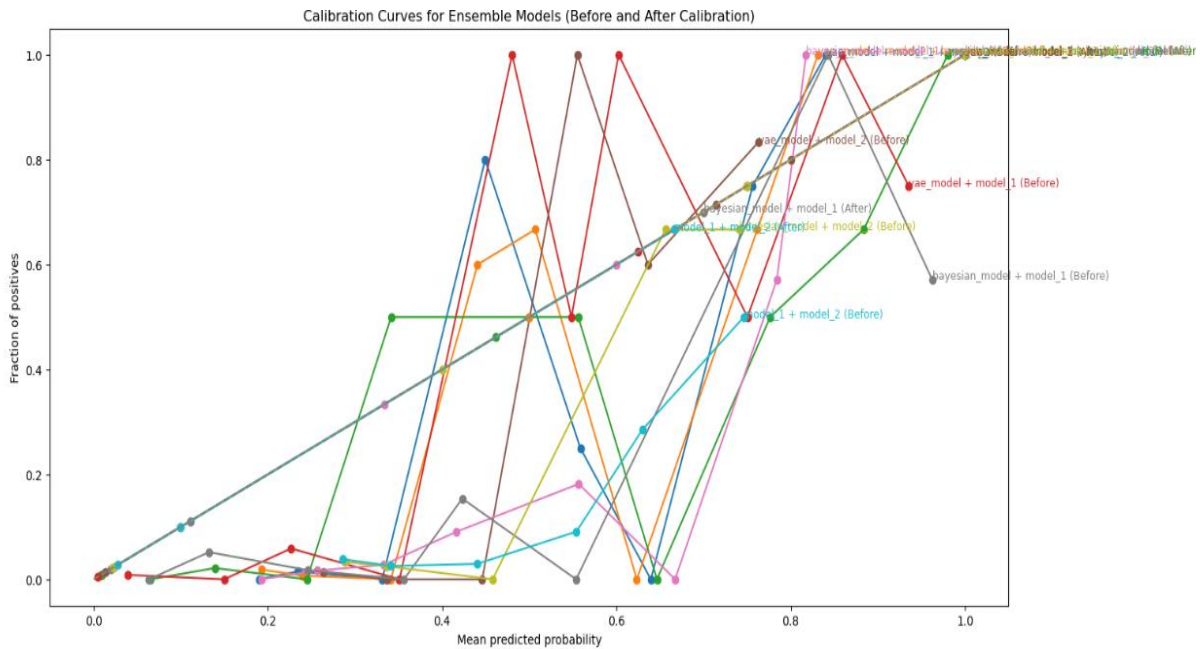
Ensemble Model	Brier Score (Before Calibration)	Brier Score (After Isotonic Calibration)	Comments
VAE + Model_2 + Model_1	0.0321	0.0254	Significant improvement after calibration.
VAE + Model_2	0.0412	0.0335	Improved accuracy after calibration.
VAE + Bayesian + Model_2 + Model_1	0.0289	0.0228	Calibration reduced overconfidence.
VAE + Bayesian + Model_2	0.0395	0.0307	Calibration improved performance.
VAE + Bayesian	0.0423	0.0330	Better post- calibration results.
Bayesian + Model_2	0.0348	0.0265	Reduced Brier score after calibration.



Ensemble Model	Brier Score (Before Calibration)	Brier Score (After Isotonic Calibration)	Comments
VAE + Model_1	0.0456	0.0361	Notable improvement with isotonic calibration.

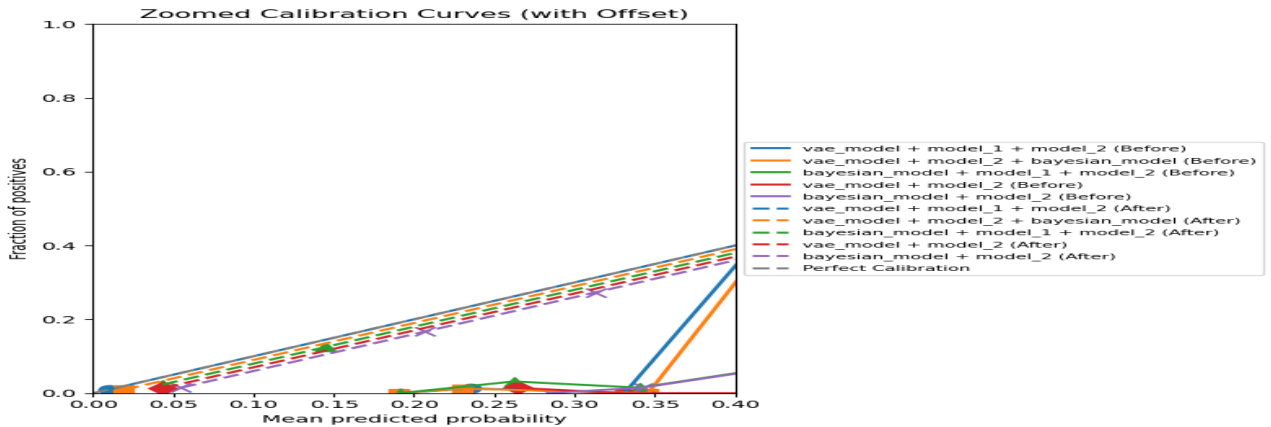


**Figure 4.4: Brier Score for Different Ensemble Models before and after Temp Scaling**



**Figure 4.5: Brier Score for Different Ensemble Models before and after Isotonic Regression**

**Zoomed Calibration**



**Figure 4.6: Brier Score for Different Ensemble Models before and after Isotonic Regression**

**Post-Calibration Insights (Suitability for Mortality Forecasting)**

**1. VAE + Flipout at Last Layer:** High entropy values,(76.81 (Global), 5.3 (Positive Class)), indicate this model has significant uncertainty in its predictions despite decent

metrics( ROC-AUC: 0.95 , Precision: 0.62, Recall: 0.77 and F1 Score: 0.69). High uncertainty is a major concern for mortality forecasting where reliable uncertainty estimation is important. Post-calibration improvements show that the model benefits from calibration. Despite decent recall, this model is **not well-suited for mortality forecasting** due to high entropy and uncertainty. The model's substantial uncertainty undermines its suitability for practical applications.

**2.VAE + Flipout All Layers (Model\_2).** This combination shows robust performance, achieves the **perfect balance** between precision (.77) and recall (.77), a highly desirable outcome in mortality forecasting, and **Platt Scaling** further enhances precision and recall, making it ideal for models with uncertain predictions. The initial score of 0.3128, significantly reduced to 0.2749 after Temperature Scaling and 0.0335 after Isotonic. However, the model has low ROC-AUC value of .90 and elevated entropy pre-calibration suggests considerable uncertainty, raising concerns about reliability in clinical settings.

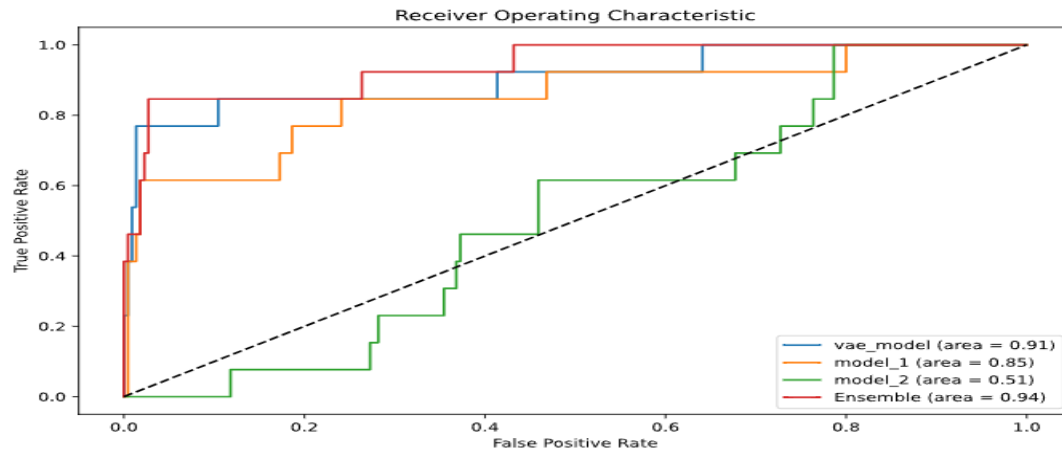
**3. Ensemble of VAE + Model\_2 + Bayesian\_Model.** This ensemble exhibits significant improvement post-calibration, reflected in the notable reduction in Brier Score especially when calibrated with **Isotonic Regression**. The high recall (.85) indicates a strong capability in identifying mortality cases. However, the elevated entropy pre-calibration suggests considerable uncertainty, raising concerns about reliability in clinical setting. Despite the improvements post-calibration, relatively lower precision (.65), indicates potential overestimation of mortality cases. Elevated entropy suggests moderate uncertainty in probability estimates. ROC\_AUC is low (.83)

**4.The ensemble of VAE + Flipout Last Layer + Bayesian** Despite the improvements post-calibration, relatively lower precision(.65) indicates potential over-estimation of mortality cases. Elevated entropy suggests moderate uncertainty in probability estimates. with the ROC-AUC around 0.86, may not be the best choices for mortality forecasting.

## 5. VAE + Model\_1 (Flipout Last Layer) + Model\_2 (Flipout All Layers) Ensemble -

The ensemble) stands out as the best performing model across all key metrics, calibration techniques, and uncertainty assessments for the following reasons. Firstly, it has **Balanced Metrics Across Precision(.65), Recall(.85), and F1 Score(.73),mean probability of positive class(.63+/-0.004)**. Post calibration the ensemble consistently achieves an **F1 score, precision, and recall of 0.77**, demonstrating a **strong balance between precision (correctly identifying positive cases) and recall (capturing as many true positives as possible)**. This is crucial in mortality forecasting, where both false positives and false negatives carry significant implications. Secondly, it has **High Accuracy and ROC-AUC** with an **ROC-AUC of 0.94**. This ensemble has proven to be highly effective at distinguishing between positive and negative classes, reinforcing its ability to perform well in mortality prediction. Thirdly **Low Entropy and Reduced uncertainty**. The ensemble exhibits **low entropy values (Global- 5.4, Positive Class- 2.5)**, indicating that the model is making **confident predictions**. This is essential in critical applications like mortality forecasting, where predictive confidence matters. Fourthly it has **Well-Calibrated Probabilities**. The ensemble demonstrates strong performance in terms of **calibration** using all three techniques (Temperature Scaling, Isotonic Regression, Platt Scaling). **Isotonic Regression** yielded a **Brier score of 0.0254**, highlighting the ensemble's reliability in providing calibrated probabilities—important for real-world deployment where uncertainty in predictions needs to be managed effectively. Fifthly ,it is **robust on Imbalanced Data**. Given the imbalanced nature of the dataset, with a small number of positive cases, the ensemble has shown robustness with **low false positives (3) and false negatives (3)**. This indicates that it can handle skewed class distributions without sacrificing performance and lastly it shows **Consistency Across Various Metrics** where unlike other ensembles that show variability across calibration techniques or suffer from

high uncertainty, this ensemble maintains **consistent performance across metrics**, making it more dependable for practical, high-stakes applications like forecasting mortality.



**Figure 4.7: ROC\_AUC Ensemble of VAE Model, Model\_1 and Model\_2**

6. **VAE + Bayesian Model** has good Pre-Calibration Metrics with an Accuracy of 0.97, Precision (0.65), Recall (0.85), F1 Score (0.73), ROC-AUC (0.89), Mean Probability of Positive Class ( $0.67 \pm 0.20$ ), Mean Probability of All Predictions ( $0.13 \pm 0.16$ ), Entropy values of 5.4 (Global), 2.8 (Positive Class) and initial Brier Score of 0.0423 (Isotonic regression). Post-Calibration Brier Score got reduced to 0.0330 (This reduction indicates that the model improves its reliability and reduces uncertainty post-calibration.). The initial high **Recall (0.85)** indicates that this model effectively identifies positive mortality cases before calibration **ROC-AUC of 0.89** pre-calibration is quite strong, showing good discrimination ability. However, the mean probability values and standard deviations show considerable uncertainty pre-calibration (e.g.,  $0.67 \pm 0.20$ ), indicating that the model is less confident in its positive class predictions. Post-calibration **temperature Scaling** leads to the most significant reduction in Brier Score and a moderate ROC-AUC of 0.78, indicating more reliable probability estimates post-calibration but slightly reduced discriminative

performance. **Isotonic Regression** increases ROC-AUC to 0.87 but results in a higher Brier Score compared to temperature scaling, suggesting less effective alignment of probabilities. **Platt Scaling** improves ROC-AUC to 0.90 but fails to reduce the Brier Score significantly, indicating potential overfitting of probability estimates. The **mean probability of positive class** shows high variability ( $0.67 \pm 0.20$ ), suggesting substantial pre-calibration uncertainty. Overall ensemble demonstrates good overall performance, with high recall and improved reliability post-calibration. It is a strong contender for mortality forecasting, especially with **Temperature Scaling**. However, due to the considerable pre-calibration uncertainty, it may not be the best standalone choice without calibration or additional model enhancements. This variability indicates that predictions for positive cases are less reliable, which is a concern for clinical applications.

7. **Bayesian model + model\_1** Model had Initial brier score of 0.2550, reduced to 0.2521 after Temperature Scaling and further to 0.0265 after Isotonic Calibration, **precision**(0.67),**Recall** ( 0.62),**F1 Score** (0.64), ROC-AUC (.88) Lower recall indicates this model misses some mortality cases, which is a drawback in clinical settings.

8. **Bayesian + Flipout All Layers (model\_2)** Model has average Metrics with Accuracy (0.96),Precision (0.67),Recall ( 0.62),F1 Score( 0.64),ROC-AUC(0.82),Mean Probability of Positive Class (  $0.55 \pm 0.16$ ), Mean Probability of All Predictions(  $0.37 \pm 0.07$ ),Entropy values of 5.43 (Global), 2.51 (Positive Class). With respect to Post-Calibration Suitability, its strength is in good precision and stable F1 Score. The mean probability values are relatively lower compared to other models, which indicates more conservative estimates. However , Lower ROC-AUC suggests weaker discriminative performance compared to other ensembles. The lower mean probability values indicate less confidence in predictions. Model as such is not as Suitable due to lower ROC-AUC and its overall performance is not optimal for mortality forecasting. With respect to Post-Calibration Suitability when

comparing Bayesian + Model\_2, Bayesian + Model\_2 has lower ROC-AUC and a wider range of probability values, indicating higher uncertainty.

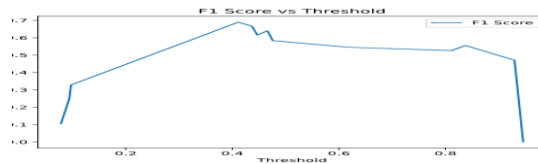
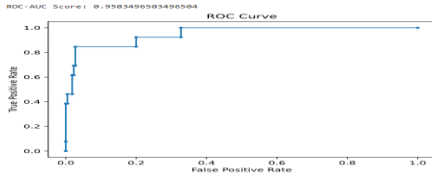
**9. Flipout All Layer + Flipout Last Layer Ensemble Analysis-**Model has good pre-calibration matrices with an Accuracy (0.96), Precision (0.67), Recall (0.62), F1 Score (0.64), ROC-AUC: (0.82), Mean Probability of Positive Class ( $0.55 \pm 0.16$ ), Mean Probability of All Predictions ( $0.37 \pm 0.07$ ), Entropy values of 5.43 (Global), 2.38 (Positive Class) and initial Brier Score of 0.3092 (Temp scaling). Post-Calibration Brier Scores was 0.2737 (Temp Scaling) with a reduction of -0.0355. Large reduction indicates that temperature scaling effectively reduces prediction errors and uncertainty. **Isotonic Regression Post-Calibration Brier Score** was 0.06 (Indicates strong improvement in alignment of predicted probabilities with true outcomes), **Platt Scaling, Post-Calibration Brier Score** was 0.94 (Indicates poor calibration and alignment, possibly due to the limited flexibility of Platt Scaling for this model). Overall ensemble had good performance across metrics and shows large reductions in Brier Score after calibration, particularly with Isotonic Regression. The pre-calibration ROC-AUC of 0.82 was lower compared to some other models, suggesting it might not have strong discriminative capability before calibration. Additionally, the mean probability of positive class shows high variability ( $0.55 \pm 0.16$ ), indicating pre-calibration uncertainty and its lower pre-calibration ROC-AUC also makes it less reliable along with low initial confidence i.

**.10. Bayesian + Flipout All Layers (Temp Scaling)** The model has Moderate performance with a tendency to underperform in recall compared to others. Although this model has relatively stable **mean probability** and **entropy**, it **doesn't perform as well** in terms of recall and calibration.

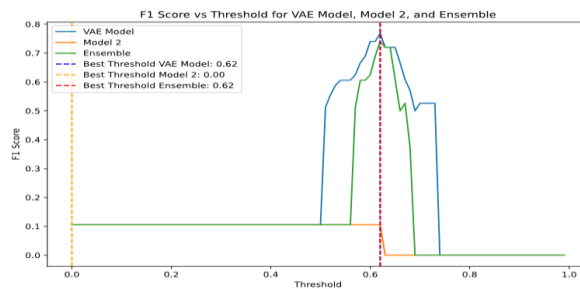
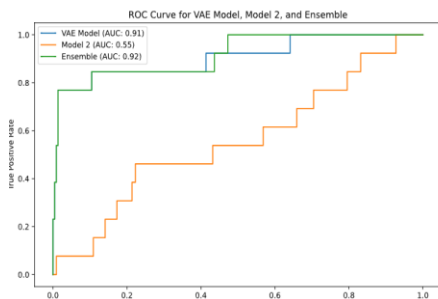
**10. VAE + Bayesian + Flipout All Layers + Last Layer** The model has good **Metrics** with ROC-AUC of 0.90, Precision of 0.65, Recall of 0.85, F1 Score of 0.73, Mean

Probability of Positive Class:  $0.63 \pm 0.05$ , Entropy values of 5.4 (Global), 2.5 (Positive Class). High ROC-AUC post-calibration indicates strong discriminative power. The model maintains a good F1 Score and recall with stable mean probability. Higher entropy in mean probability indicates some variability, suggesting the need for further adjustments. The **model** is suitable due to its balance of metrics and ROC-AUC improvement post-calibration.

VAE+ Model\_1

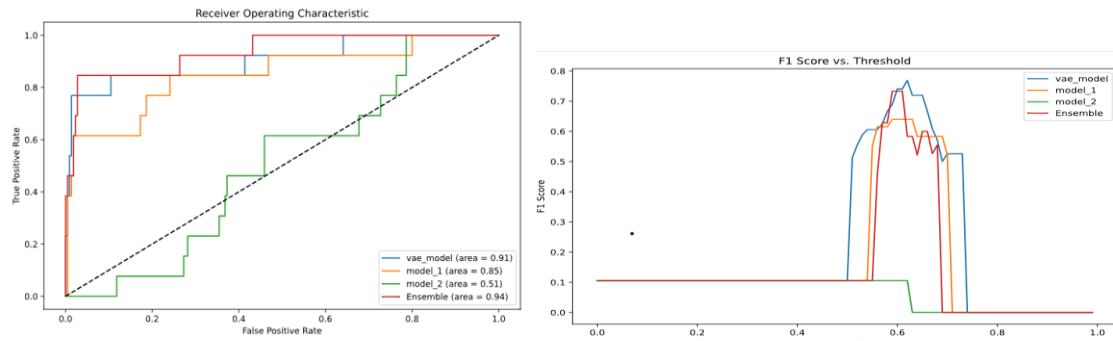


VAE+ Model\_2



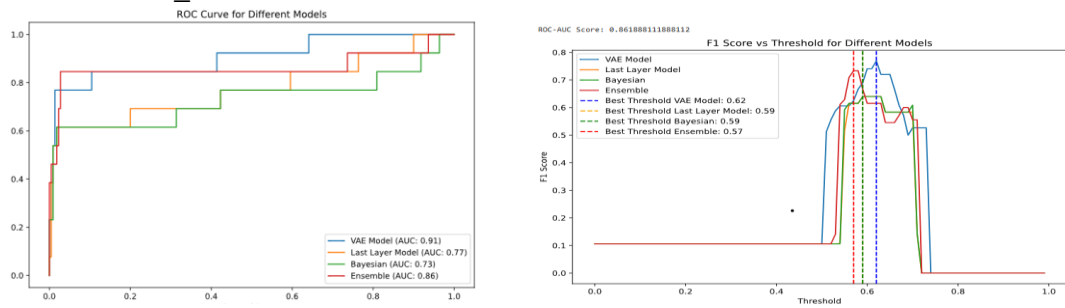
VAE+Model\_1+Model\_2



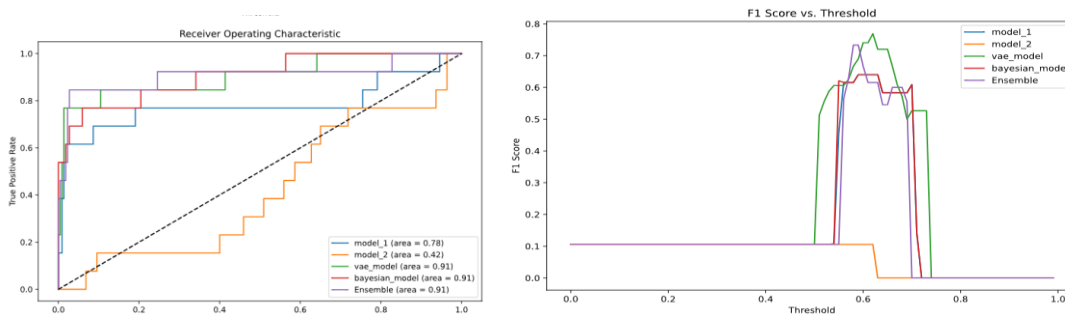


**Figure 4.8 Comparison of ROC of Ensemble Models I**

**VAE+Model\_1 +BNN**

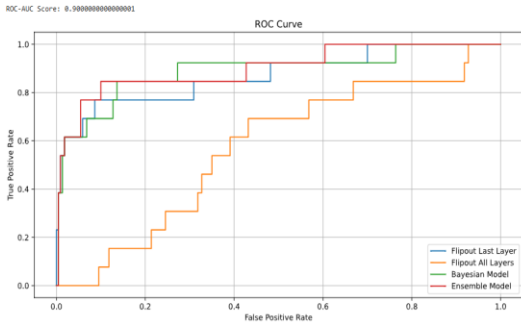


**VAE+ Model\_1+ Model\_2 +BNN**

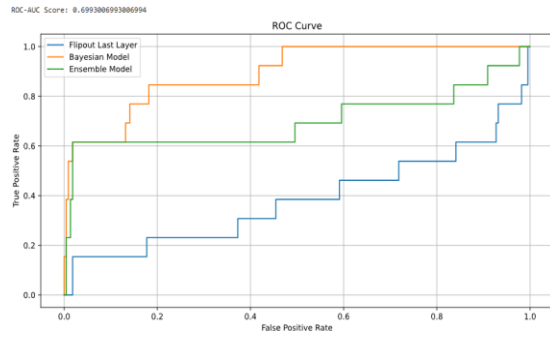


**Figure 4.9 Comparison of ROC of Ensemble Models II**

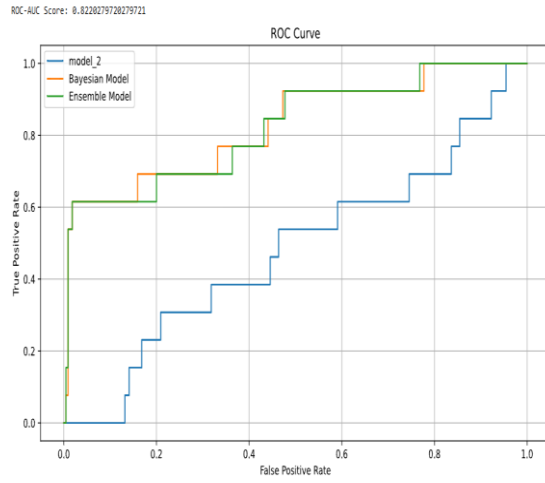
Model\_1+Model\_2+BNN Model



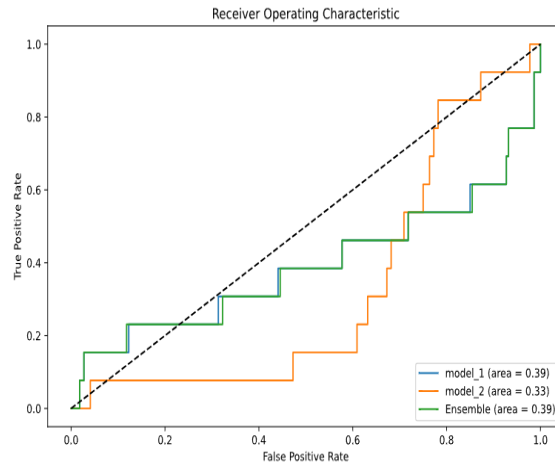
Model\_1 + BNN Model



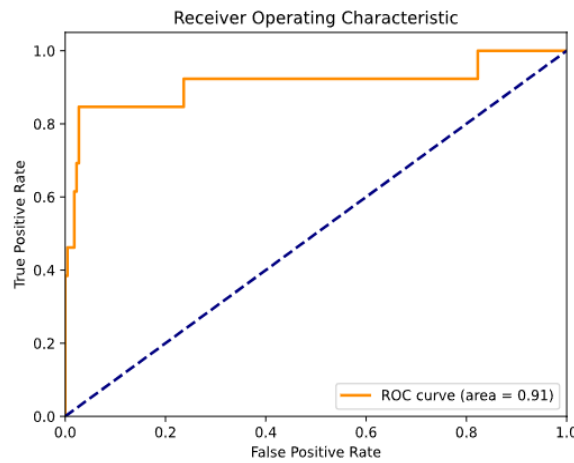
Model\_2+BNN Model



Model\_1+Model\_2



### VAE +Bayesian Model



**Figure 4.10 Comparison of ROC of Ensemble Models III**

The **ensemble of VAE + Model\_1 (Flipout Last Layer) + Model\_2 (Flipout All Layers)** is recommended as the **best ensemble for mortality forecasting** due to its **High precision and recall**, ensuring that both true positives and true negatives are captured effectively **Low uncertainty and well-calibrated outputs**, making it suitable for real-world deployment in healthcare settings, **Consistency in performance** across various calibration techniques, further enhancing its reliability in uncertain environments, **Ability to handle**

**imbalanced data** with minimal false positives and false negatives, which is crucial in clinical applications where the stakes are high.

This ensemble balances performance and uncertainty management, making it the most suitable for mortality forecasting tasks in your research.

**Best Overall Performance -VAE + Flipout Last Layer + All Layers** offered the highest **ROC-AUC score** and maintained strong precision, recall, and F1 score. This made it an excellent choice before calibration.

**Good Balance-VAE + Flipout All Layers** and **VAE + Flipout Last Layer** showed excellent balance between precision, recall, and AUC scores. These models were well-suited for scenarios where both metrics needed to be balanced. **Bayesian + Flipout Last Layer** had higher **precision** at the expense of recall, making it suitable when reducing false positives was critical. Meanwhile, **VAE + Flipout Last Layer** had high recall with slightly lower precision, which was good for identifying more positive cases.

### 4.3 Research Question Three

Whether the variable of importance derived from model explainability techniques of Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations, commonly known as SHAP can explain and interpret the model for prognostication of postoperative mortality with identical results?

Two techniques of model Interpretability were experimented. Understanding of the feature importance can help validate the model's behaviour, identify potential biases, and improve model transparency and interpretability. In LIME (Local Interpretable Model-agnostic Explanations), the prediction of an outcome depends on the presence or absence of a combination of features which through linear regression decide the outcome. For example, the feature value of  $\leq 0.00$  may indicate a threshold condition. If the value of the "Sepsis" feature, first feature in a given instance is less than or equal to 0.00, then the value  $\leq 0.00$  might indicate "no sepsis." The weight of feature which is highest for all variables will shows that the feature has a significant impact on the model's prediction. Thus, changes in this feature's value can meaningfully affect the prediction.

The other model interpretability technique is SHAP. In SHAP the expected value, also known as the base value, represents the model's average prediction for all data points in the dataset. It serves as a reference point for interpreting SHAP values. When interpreting SHAP values, the expected value provides context for understanding the contribution of each feature to the model's prediction. SHAP values represent the difference between the actual prediction for a specific instance and the expected value.

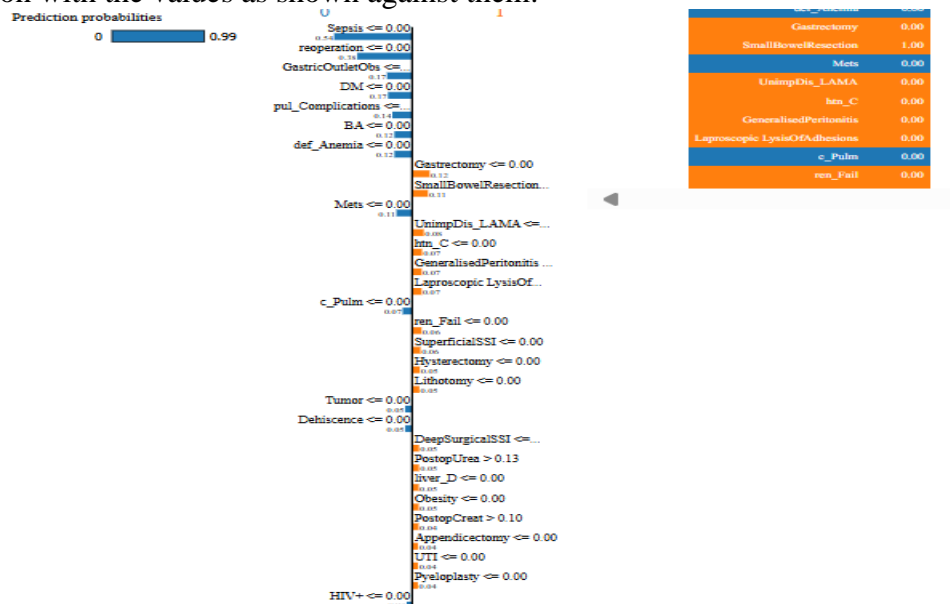
At the Local Level two single identical data instances were chosen and results of **LIME Coefficients** and SHAP values were evaluated to understand the features contribution for comparison . Both Lime and SHAP identified some common features with high feature importance as under: -.

In one data instance. a 10<sup>th</sup> row of test data, X\_test, the feature names and features values were as under: -

Reoperation 0.00, Readmission 0.00, ChroLiverDis 0.00, Tumor 0.00, renal\_Failure 0.00, PostopUrea 0.22, UnimpDis\_LAMA 0.00, Hysterectomy 0.00, Mets 0.00, hypo\_Thyroidism 0.00, Generalised Peritonitis 0.00 cardiac\_Complication 0.00,

liver\_Disease 0.00, Laproscopic LysisOfAdhesions 0.00, Oesophagotomy 0.00, SmallBowelResection 1.00,Lithotomy 0.00, GastricOutletObs 0.00, Hernioplasty 0.00, LapCholi 0.00, SuperficialSurgicalSiteInfection 0.00, hypertension\_Chronic 0.00, Gastrectomy 0.00, HIV+ 0.00, OpenCholi 0.00, Obesity 0.00 Dehiscence 0.00 ,UTI 0.00, Omentoplasty 0.00 etc.

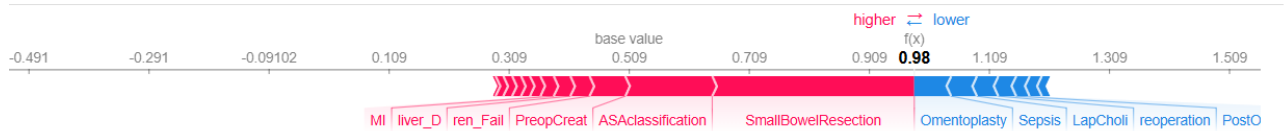
Lime drew the following table with a probability of .99 for positive outcome of mortality. As the values suggest, and table confirms that small bowel resection with a coefficient of .11 and postop urea with a coefficient of .09 predicted mortality. Sepsis, reoperation, Readmission, ChroLiverDisease, Tumor, and ren\_Failure are on the blue side, indicating that they are below the threshold and are considered to be negatively impacting the prediction with the values as shown against them.



**Figure:4.11 Data instance 10th row with interpretation from LIME**

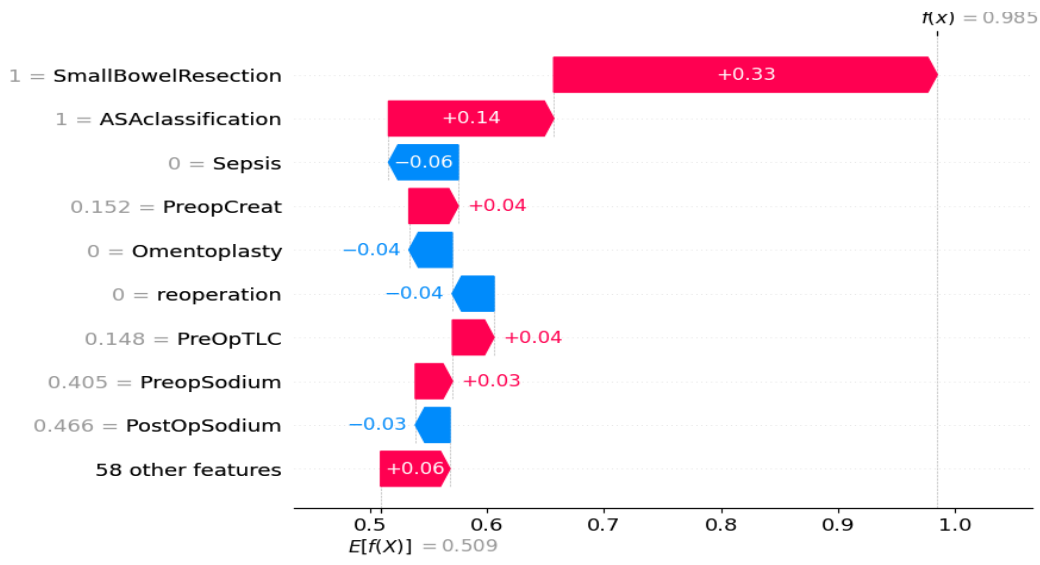
For the same data instance , SHAP too listed Small Bowel resection with the highest SHAP value of .33 .however, other features contributing to motility are different which include

ASA classification with a SHAP Value of .13 ,PreopCretinine with SHAP value of .04, renal\_Failure, 0.01, liver Disease, 0.02 contributing to mortality. SHAP value of 0 means that the absence or presence of the feature has a neutral impact on the model's prediction compared to the expected value whereas Model output (fx) of 1 means that prediction for this instance is positive. The red color typically represents high feature values or feature effects that contribute positively to the prediction.



The SHAP values are Small Bowel Resection,0.33,ASAclassification, 0.14, PreopCreatinine, 0.04, Pre OP TLC ,.04, PreopSodium .03, renal\_Failure, 0.01, liver Disease, 0.02

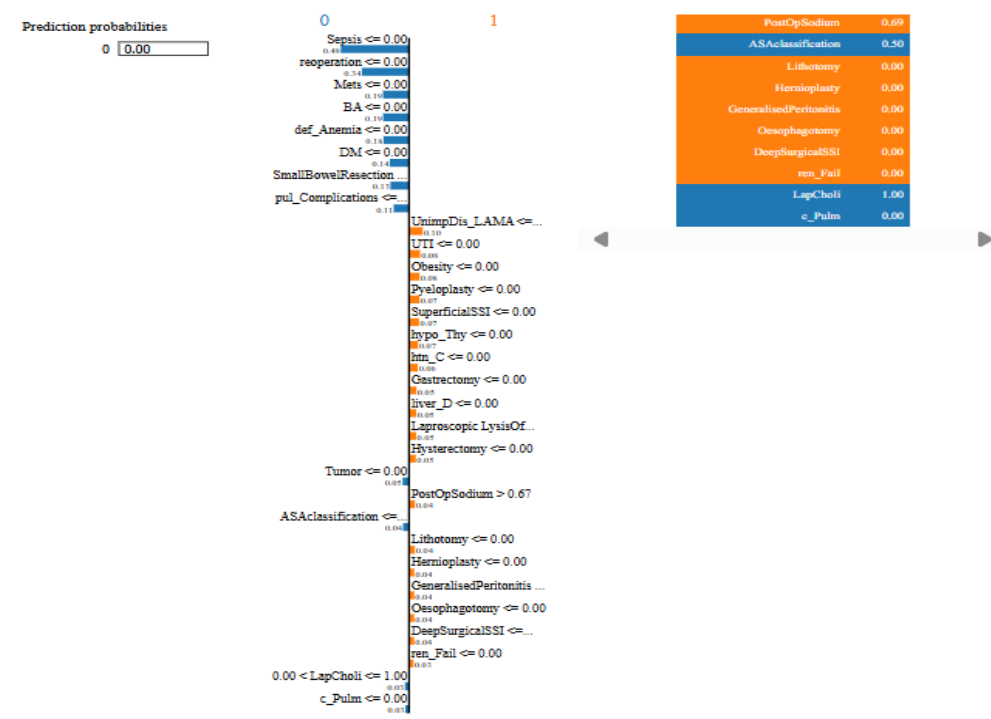
**Figure:4.12 Data Instance 10<sup>th</sup> with interpretation from SHAP**



**Figure:4.13 SHAP Values of Important Variables**

In another data instance of the 6th row of test data where surgery was done was Lap Cholecystectomy, while both LIME and SHAP predicted no Mortality, Feature values and LIME coefficients are as under :-

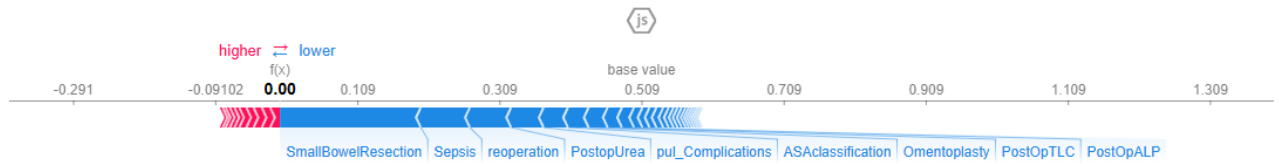




**Figure:4.14 Data Instance 6<sup>th</sup> with interpretation from LIME**

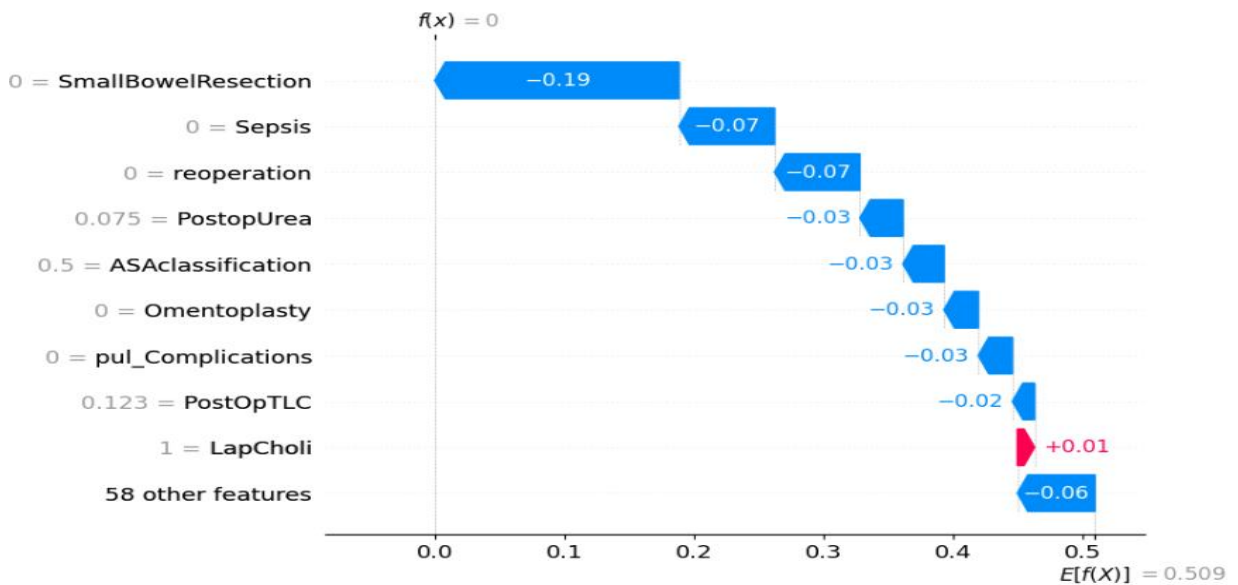
LapCholi had a coefficient value of 0.03 and is highlighted in blue, indicating that its presence reduces the probability of class 1. Despite its positive value, LapCholi seems to have a negligible effect on the overall prediction, given its low coefficient value compared to other features. Therefore, the model's high confidence in predicting class 0 is primarily driven by the absence or low influence of features associated with class 1, as indicated by their low or negative coefficient values. In the case of SHAP too for the same data instance has features which lowers the risk of death. The SHAP force plot has a base value is 0.509 and the color of the features is blue except for a bar where few variables LapCholi with SHAP Value of 0.011, PreOpTLC, with SHAP value of 0.009, Renal Failure, with SHAP value of 0.008, Organ Space SSI, with SHAP Value: 0.005 etc are gathered and have negligible effect over the model. . In the context of i.e. a SHAP force plot, blue color indicates that the feature values are lower than the corresponding baseline values. Blue

features contribute to pushing the model's prediction towards a lower output compared to the baseline prediction.



**Figure:4.15 Data Instance 6<sup>th</sup> with interpretation from SHAP**

Shap waterfall graph displays the values of coefficients for features and similar to force plot.



**Figure:4.16 Data Instance 6<sup>th</sup> with interpretation from SHAP Showing SHAP Values in Waterfall**

Summary of feature Importance at the Global level however varied for LIME and SHAP with different Coefficients and weights respectively as also the order for feature importance for predicting mortality.

In LIME output, the coefficients are used to interpret the contribution of each feature towards the prediction made by the model for a particular instance. When a coefficient is displayed in orange and it is greater than 0, it indicates a positive contribution of that feature towards the prediction of the corresponding class label. Conversely, when the coefficient is less than 0, it indicates a negative contribution. Here's how to interpret the coefficients in orange:

Greater than 0: This means that an increase in the value of the feature will likely result in an increase in the probability of the corresponding class label being predicted.

Less than 0: This means that an increase in the value of the feature will likely result in a decrease in the probability of the corresponding class label being predicted.

While LIME's predictions are Local by nature a consolidated summary of mean lime coefficients was calculated, the value of coefficients is as under:

Less than 0: This means that an increase in the value of the feature will likely result in a decrease in the probability of the corresponding class label being predicted.

While LIME's predictions are Local by nature a consolidated summary of mean lime coefficients was calculated ,the value of coefficients is as under:

**Table 4.9 LIME OUTPUT WITH VALUES OF COEFFICIENTS**

Sr No	Feature	Positive Coefficient	Sr No	Feature	Negative Coefficient
0	Sepsis > 0.00	2.327883	62	PostOpALP .11<=.14	-0.002194
1	UnimpDis_LAMA <= 0.00	1.250733	63	PostOpSGPT <=.01	-0.008995
2	Obesity <=0	1.1906	64	PreopPotassium <=.12	-0.009662
3	Hysterectomy <= 0.00	1.174704	65	PreOpBilD <=.02	-0.010606
4	Lithotomy <= 0.00	1.143281	66	PreopSodium .62<=.68	-0.01169
5	ren_Fail <= 0.00	1.113642	67	PostopCreat <=.10	-0.012598
6	PostopUrea > 0.13	1.106336	68	PostOpPotassium <=.48	-0.013398
7	GeneralisedPeritonitis <= 0.00	1.100541	69	PostopCreat <=.06	-0.015088
8	c_Complication <= 0.00	1.092594	70	PreopUrea<=.07	-0.015109
9	liver_D <= 0.00	1.081784	71	PreopCreat .03<=.04	-0.015245
10	SmallBowelResection > 0.00	1.054477	72	PreopCreat <=.03	-0.018905
11	Laprosopic_LysisOfAdhesions <= 0.00	1.036002	73	PostOpSGPT <=.02	-0.024968
12	Oesophagotomy <= 0.00	1.034304	74	PostopUrea <=.10	-0.027095
13	MRM <= 0.00	1.003282	75	PostopUrea .10 <.=13	-0.030578
14	Hernioplasty <= 0.00	0.993488	76	PreopUrea .07 <.=13	-0.031156

15	LapCholi <= 0.00	0.97768	77	PostOpALP >.14	-0.031661
16	GastricOutletObs <= 0.00	0.938631	78	PostopUrea <=.07	-0.040938
17	SuperficialSSI <= 0.00	0.908357	79	PostOpPotassium >.48	-0.042257
18	UTI <= 0.00	0.832508	80	PostOpSGPT <.01	-0.043488
19	htn_C <= 0.00	0.799847	81	htn_C >0.00	-0.05797
20	HIV <= 0.00	0.784619	82	OpenCholi >0.00	-0.062354
21	OpenCholi <= 0.00	0.777723	83	PostOpBilT >.08	-0.065091
22	c_Pulm <= 0.00	0.757509	84	pul_Complications <=0	-0.070766
23	Gastrectomy <= 0.00	0.709402	85	Hysterectomy >0.00	-0.087939
24	BA <= 0.00	0.689688	86	Herniotomy <=0	-0.088217
25	Dehiscence <= 0.00	0.67251	87	liver_D >.00	-0.089892
26	DeepSurgicalSSI <= 0.00	0.625793	88	PreOpBilT >.04	-0.090484
27	DiagLaprot <= 0.00	0.548667	89	PostOpBilD >.04	-0.113323
28	ASAclassification 0.50 <= 1.09	0.505842	90	DiagLaprot >0.00	-0.119151
29	PostOpSGPT >0.02	0.378935	91	PostOpSodium <.53	-0.128315
30	Pyeloplasty <=0.00	0.37154	92	PreOpSGOT >.04	-0.136333
31	ChroLiverDis >0.00	0.336464	93	ASAclassification <= .50	-0.145592
32	PreopCreat >	0.328804	94	PreOpTLC >.19	-0.154272
33	PostopCreat >0.06	0.311245	95	DM<=.00	-0.174515
34	Tumor >0	0.302301	96	ALP >.15	-0.205158

35	PostOpPotassium <=.35	0.299689	97	DeepSurgicalSSI >.00	-0.222295
36	Reoperation >0.00	0.253629	98	ren_Fail >0.00	-0.222651
37	PostOpSodium >0.67	0.219148	99	SmallBowelResection <=.00	-0.295664
38	PreOpTLC <=0.11	0.201695	100	Omentoplasty <=0.00	-0.452172
39	PreopUrea >.14	0.195143	101	hypo_Thy <=0.00	-0.712808
40	Mets >0.00	0.165948	102	Mets <0.00	-1.475026
41	PostOpSGOT >.05	0.110666	103	ChroLiverDis <=0.00	-1.575212
42	Appendicectomy <=0.00	0.109825	104	Readm <=0.00	-2.629983
43	Omentoplasty >.00	0.087355	105	Sepsis <= 0.00	-2.645693
44	ALP <=.09	0.072571	106	Tumor <=0.00	-3.047706
45	PreOpTLC <=.15	0.059122	107	Reoperation <=0.00	-3.353869
46	PreOpSGOT <=.02	0.056368			
47	def_Anemia <=0.00	0.046376			
48	DM >.00	0.039826			
49	PreOpBilT .01<=.02	0.037867			
50	pul_Complications >.00	0.03543			
51	PreopSodium <=.51	0.03386			
52	ALP .09<=.12	0.033165			
53	PostOpALP <=.08	0.033138			
54	PostOpBilT .04<=.05	0.018208			
55	PreOpSGOT .02<-<=.02	0.01535			

56	PostOpTLC <+.17	0.011931
57	PostOpPotassium <=.42	0.011038
58	PostOpBilD .02<=.04	0.010766
59	PostOpBilD <=.02	0.010325
60	ALP .12<=.15	0.007825
61	PreOpBilD <=.01	0.004383

Positive Coefficients indicate that the presence or higher value of a feature increases the likelihood of the predicted outcome whereas Negative Coefficients indicate that the presence or higher value of a feature decreases the likelihood of the predicted outcome. The coefficients indicate the magnitude of the feature's impact on the outcome. For instance, a higher positive coefficient means the feature strongly increases the likelihood of the outcome, while a higher negative coefficient means the feature strongly decreases the likelihood of the outcome. In this ,important features identified are as under:

For example ,Sepsis > 0.00: 2.32. indicates that the presence of sepsis significantly increases the likelihood of the predicted outcome, with a high coefficient of 2.327.

UnimpDis\_LAMA <= 0.00: 1.2. If the patient is nor leaving against medical advice (LAMA), it increases the likelihood of the outcome of mortality .

Obesity <= 0.00: 1.19. Absence of obesity increases the likelihood of the predicted outcome. A counterintuitive result but features are only suggestive base on correlation and are not considered causal

Hysterectomy  $\leq 0.00$ : 1.17. Not having had a hysterectomy increases the likelihood of the outcome as it is simple procedure.

Lithotomy  $\leq 0.00$ : 1.14. Not having undergone a lithotomy (surgical removal of stones) increases the likelihood of the outcome as it is a simple procedure.

ren\_Fail  $\leq 0.00$ : 1.113642190240061. Absence of renal failure increases the likelihood of the outcome. This is Acute Renal injury following Surgery which gets treated subsequent to surgery over the time in many cases. Patients with Acute renal injury might be receiving more intensive monitoring and aggressive treatment, leading to better-than-expected outcomes.

PostopUrea  $> 0.13$ : 1.10. Postoperative urea levels above 0.13 increase the likelihood of the outcome. This is because of the fact that patient with Chronic renal disease and those with Acute Kidney injury not responding to management will have high Blood urea level and has adverse prognosis.

Other positive features follow similar logic, where the absence of a condition or the presence of a certain threshold value increases the likelihood of the predicted outcome.

Many of the results appear counterintuitive but these coefficients only tell about correlations and not causation.

With respect to SHAP the red colour denotes high value and blue colour low value on either side positive or negative. Positive and Negative weights are as under:



**Table 4.10 SORTED POSITIVE SHAP VALUES FOR VARIABLES**

Sorted positive SHAP values for y=1:	
SmallBowelResection	0.18234673637678156
Sepsis	0.09819226988202501
PostOpSGPT	0.046634751852670187
ChroLiverDis	0.030318443970439337
Hernioplasty	0.02781855255131928
ASAclassification	0.02287675129673326
PostopCreat	0.02252254410640177
pul_Complications	0.021551255015035367
DM	0.017436566519715334
DeepSurgicalSSI	0.011857645745236536
PostOpPotassium	0.011688300348512788
PostOpSodium	0.011521893848428631
liver_D	0.010122196049912305
PostopUrea	0.009780539221517866
OrganSpaceSSI	0.008928766862425412
PreOpSGOT	0.008894080019202343
PreOpBilD	0.007439960104409799
PreopCreat	0.006710931263822279
PostOpTLC	0.004825369677437139
PostOpSGOT	0.0047638777479090495
SuperficialSSI	0.004409163624958699
OpenCholi	0.0041424415158161955
Lithotomy	0.002906090893402156
Mets	0.0024765917624669648
GeneralisedPeritonitis	0.002240953234739925
Herniotomy	0.001876411917357886
Oesophagotomy	0.001790581359688539
UnimpDis_LAMA	0.00175027051553751
GastricOutletObs	0.0015413009987989396
PreOpBilT	0.0013361866138165658
MRM	0.0011760715347959709

Sorted negative SHAP values for y=1:	
ren_Fail	-0.014429419335756729
ALP	-0.013658995537857866
Omentoplasty	-0.011305369345782048
PostOpBilD	-0.009119278806813486
reoperation	-0.008823615329846794
PreOpTLC	-0.0062409372752514685
PostOpBilT	-0.0060433165160105546
DiagLaprot	-0.004741266119464331
PreopSodium	-0.004716811213989488
PreopUrea	-0.004567400681978511
LapCholi	-0.004551492957978039
Tumor	-0.0031879373855149256
CVA	-0.0031399630739585124
Readm	-0.002738235088495108
c_Pulm	-0.002682942485646804
htn_C	-0.0023062002168726167
Prostectomy	-0.0015687732788280573
PostOpALP	-0.0014664571225033324
BA	-0.0010370133546462339
Laprosopic LysisOfAdhesions	-0.0010214386162803268
MI	-0.0008698971466552228
Hysterectomy	0.0007675449054255132
c_Complication	-0.0006451518190264177
Hemiplegia	-0.0005717984768389342
Obesity	-0.0005436729122828631
Dehiscence	0.0005253967634293204
Nephrectomy	-0.00024674727294986
PreopPotassium	-4.9570656069694016e-05
hypo_Thy	-3.13119389807454e-05

Appendicectomy	0.0008994733931824052
UTI	0.0008753648437528089
def_Anemia	0.0006100325454171744
Pyeloplasty	0.0002463559002564049

Meaning of Positive SHAP Values for features with positive SHAP values:

Small Bowel Resection (0.182): Indicates that having Small Bowel Resection significantly increases the likelihood of the outcome ( $y=1$ ).

Sepsis (0.098): Presence of Sepsis increases the likelihood of the outcome.

Post Op SGPT (0.0466): Higher Postoperative SGPT levels increase the likelihood of the outcome.

Chronic Liver Dis (0.0303): Chronic Liver Disease increases the likelihood of the outcome.

Hernioplasty (0.0278): Undergoing Hernioplasty increases the likelihood of the outcome.

ASA classification (0.0229): Higher ASA classification (indicating worse preoperative health) increases the likelihood of the outcome.

Post op Creatinine (0.0225): Higher postoperative creatinine levels due to Chronic Renal Failure /Acute Kidney injury increase the likelihood of the outcome.

Pulmonary Complications (0.0216): Pulmonary complications increase the likelihood of the outcome.

DM (0.0174): Having Diabetes Mellitus increases the likelihood of the outcome.

Deep Surgical SSI (0.0119): Deep Surgical Site Infections increase the likelihood of the outcome.

Post Op Potassium (0.0117): Higher postoperative potassium levels increase the likelihood of the outcome.

Post Op Sodium (0.0115): Higher postoperative sodium levels increase the likelihood of the outcome.

liver\_D (0.0101): Liver disease increases the likelihood of the outcome.

PostopUrea (0.0098): Higher postoperative urea levels increase the likelihood of the outcome.

OrganSpaceSSI (0.0089): Organ/space surgical site infections increase the likelihood of the outcome.

Pre Op SGOT (0.0089): Higher preoperative SGOT levels increase the likelihood of the outcome.

Pre Op BilD (0.0074): Higher preoperative direct bilirubin levels increase the likelihood of the outcome.

Preop Creat (0.0067): Higher preoperative creatinine levels increase the likelihood of the outcome.

PostOpTLC (0.0048): Higher postoperative total leukocyte count increases the likelihood of the outcome.

PostOpSGOT (0.0048): Higher postoperative SGOT levels increase the likelihood of the outcome.

SuperficialSSI (0.0044): Superficial surgical site infections increase the likelihood of the outcome.

OpenCholi (0.0041): Undergoing open cholecystectomy increases the likelihood of the outcome.

Lithotomy (0.0029): Undergoing lithotomy increases the likelihood of the outcome.

Mets (0.0025): Presence of metastasis increases the likelihood of the outcome.

Generalised Peritonitis (0.0022): Generalized peritonitis increases the likelihood of the outcome.

Herniotomy (0.0019): Undergoing herniotomy increases the likelihood of the outcome.

Oesophagotomy (0.0018): Undergoing oesophagotomy increases the likelihood of the outcome.

UnimpDis\_LAMA (0.0018): Discharge against medical advice increases the likelihood of the outcome for patients in the system compared to those who have left against medical advice.

GastricOutletObs (0.0015): Gastric outlet obstruction increases the likelihood of the outcome.

PreOpBilT (0.0013): Higher preoperative total bilirubin levels increase the likelihood of the outcome.

MRM (0.0012): Undergoing modified radical mastectomy increases the likelihood of the outcome.

Appendectomy (0.0009): Undergoing appendectomy increases the likelihood of the outcome.

UTI (0.0009): Urinary tract infections increase the likelihood of the outcome.

def\_Anemia (0.0006): Deficiency anaemia increases the likelihood of the outcome.

Pyeloplasty (0.0002): Undergoing pyeloplasty increases the likelihood of the outcome.

Meaning of Negative SHAP Values for features with negative SHAP values:

ren\_Fail (-0.0144): Renal failure decreases the likelihood of the outcome (y=1).Pertains to Acute Kidney injury following Surgery

ALP (-0.0137): Higher alkaline phosphatase levels decrease the likelihood of the outcome.

Omentoplasty (-0.0113): Undergoing omentoplasty decreases the likelihood of the outcome.

PostOpBilD (-0.0091): Higher postoperative direct bilirubin levels decrease the likelihood of the outcome.

reoperation (-0.0088): Need for reoperation decreases the likelihood of the outcome.

PreOpTLC (-0.0062): Higher preoperative total leukocyte count decreases the likelihood of the outcome.

PostOpBilT (-0.0060): Higher postoperative total bilirubin levels decrease the likelihood of the outcome.

DiagLaprot (-0.0047): Diagnostic laparoscopy decreases the likelihood of the outcome.

PreopSodium (-0.0047): Higher preoperative sodium levels decrease the likelihood of the outcome.

PreopUrea (-0.0046): Higher preoperative urea levels decrease the likelihood of the outcome.

LapCholi (-0.0046): Undergoing laparoscopic cholecystectomy decreases the likelihood of the outcome.

Tumor (-0.0032): Presence of a tumor decreases the likelihood of the outcome. A Benign tumor

CVA (-0.0031): Cerebrovascular accident decreases the likelihood of the outcome.

Readm (-0.0027): Readmission decreases the likelihood of the outcome.

c\_Pulm (-0.0027): Chronic pulmonary disease decreases the likelihood of the outcome.

htn\_C (-0.0023): Controlled hypertension decreases the likelihood of the outcome.

Prostectomy (-0.0016): Undergoing prostatectomy decreases the likelihood of the outcome.

PostOpALP (-0.0015): Higher postoperative alkaline phosphatase levels decrease the likelihood of the outcome.

BA (-0.0010): Bronchial Asthma decreases the likelihood of the outcome.

Laparoscopic LysisOfAdhesions (-0.0010): Laparoscopic lysis of adhesions decreases the likelihood of the outcome.

MI (-0.0009): Myocardial infarction decreases the likelihood of the outcome.

Hysterectomy (-0.0008): Undergoing hysterectomy decreases the likelihood of the outcome.

c\_Complication (-0.0006): Cardiac Complications decrease the likelihood of the outcome.

Hemiplegia (-0.0006): Hemiplegia decreases the likelihood of the outcome.

Obesity (-0.0005): Obesity decreases the likelihood of the outcome.

Dehiscence (-0.0005): Wound dehiscence decreases the likelihood of the outcome.

Nephrectomy (-0.0002): Undergoing nephrectomy decreases the likelihood of the outcome.

PreopPotassium (-0.00005): Higher preoperative potassium levels decrease the likelihood of the outcome.

hypo\_Thy (-0.00003): Hypothyroidism decreases the likelihood of the outcome

While there is a dissimilarity in how the variables weights are calculated in LIME and SHAP ,there is considerable amount of match between variables of importance in two methods of interpretability. Both LIME and SHAP has identified similar features impacting the mortality in surgical procedures .Some of the features negatively impacting the outcome are shared by both for example obesity <0 with coefficient 1.1906 in LIME and -0.0005436729122828631 in SHAP, Alkaline Phosphate 0.11 to <= 0.14 with coefficient -0.002194086349926584 and -0.0137 in SHAP. Omentoplasty feature with value <=0 has a LIME coefficient of -.45 whereas SHAP has a weight of -0.011,Postop Direct

Bilirubin with value with  $>.04$  has a LIME coefficient of  $-0.11$  and a SHAP weight of  $.009$ , PreopTLC  $<.11$  has a LIME coefficient of  $.20$  whereas PreopTLC  $>.19$  has a negative LIME coefficient of  $.17$  whereas SHAP weight for the feature is  $-.006$ . The same is true for other variables like PostopBilirubin Total (LIME if  $>.08$ , a coefficient of  $-.06$  and SHAP weight of  $-.006$ , Diagnostic Laprotomy  $>0$ , a LIME weight of  $-0.11$  and SHAP weight of  $-.004$ , PreopSodium  $.62 \leq .68$ , a LIME Coefficient of  $.011$  and SHAP weight of  $-.004$ , PreopUrea  $\leq .07$ , a LIME coefficient of  $-.015$  and a SHAP weight of  $-.004$ . Lap Choli  $\leq 0$  has a LIME coefficient of  $.93$  whereas SHAP has a weight of  $-.004$  meaning the same thing. For patients who did not undergo Laparoscopic Cholecystectomy, the LIME model assigns a coefficient of  $0.97$ . This suggests that not having the surgery is associated with a higher risk of the outcome (mortality) whereas SHAP weight is  $-.0047$ . For feature Chronic Pulmonary disease  $<0$ , LIME coefficient is  $.75$  whereas SHAP value is  $-.002$ .

There are however a few differences. While LIME emphasizes a more local perspective, focusing on individual predictions, SHAP provides a global understanding of feature contributions across all predictions. For example, LIME identifies Sepsis as the most important variable with a contribution of  $2.32$ , whereas SHAP assigns it a mean value of  $0.09$ . This discrepancy highlights the different interpretations and utility of each method. Reoperation  $>0$  has a Lime coefficient of  $.25$  whereas SHAP has a negative weight of  $-.008$ . In case of Tumor  $>0$ , LIME assigns a value of  $.30$  whereas SHAP has a negative weight of  $-.003$ . Omentoplasty  $>0$   $.087$  and SHAP has a feature weight of  $-.011$ .

Here's a simplified and clearer analysis of the variables' importance according to both LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) values. The first 10 variables in descending order as per the LIME and SHAP are as under :

For LIME Importance Scores (Descending Order) include

Sepsis > 0.00: 2.32

Postoperative Urea > 0.13: 1.10

Small Bowel Resection > 0.00: 1.05

ASA Classification 0.50 to <= 1.00: 0.50

Postoperative SGPT > 0.02: 0.37

Chronic Liver Disease > 0.00: 0.33

Preoperative Creatinine > 0.06: 0.32

Postoperative Creatinine > 0.10: 0.31

Tumor > 0.00: 0.30

Reoperation > 0.00: 0.25

These values indicate how much each variable contributes positively to the likelihood of the outcome, with higher values reflecting greater influence on the model's predictions (Ribeiro et al., 2016). Gonzalez and Franks (2018) highlight the local interpretability aspect of LIME, emphasizing its utility in understanding specific predictions. LIME Values show that Sepsis is the most important variable with a significant positive impact (2.32), followed by Postoperative Urea and Small Bowel Resection. These variables contribute strongly to the model's predictions.

In contrast, the same 10 variables assessed by SHAP provide different mean values for importance:

Mean SHAP Values for the Same Variables include

Sepsis: 0.09

Postoperative Urea: 0.009

Small Bowel Resection: 0.18

ASA Classification: 0.02



Postoperative SGPT: 0.04

Chronic Liver Disease: 0.03

Preoperative Creatinine: 0.006

Postoperative Creatinine: 0.002

Tumor: -0.003

Reoperation: -0.008

SHAP values reflect the average contribution of each feature to the prediction across all instances, providing a consistent measure of feature importance. Notably, while some variables, like Sepsis and Small Bowel Resection, retain a positive importance in both methods, others show significant differences in their contributions (Lundberg & Lee, 2017; Molnar, 2020). SHAP Values approach is a more nuanced view, indicating that while Sepsis is still important, its mean contribution is much lower (0.09). The values for most other variables are considerably lower, with some like Reoperation having negative SHAP values, suggesting they might decrease the risk or likelihood of the outcome being predicted.

This comparison highlights the differences in how LIME and SHAP assess variable importance, with LIME often indicating stronger effects overall compared to the more granular, average contributions shown by SHAP.

Choosing between LIME and SHAP depends on the specific needs and context, as each method has its strengths and weaknesses. The considerations for use of LIME decision are:

LIME is to be used When:

Local Interpretability: When quick insights into individual predictions is required rather than global feature importance.

**Simplicity:** A straightforward method that provides easy-to-understand explanations without deep statistical grounding is needed

**Speed:** A faster results is need of the hour, as LIME can be computationally less intensive compared to SHAP, especially with large datasets.

**SHAP is to be used When:**

**Consistency and Fairness** have prime importance as SHAP values are based on game theory and provide a consistent approach to feature importance, ensuring that the contributions of features add up to the model's output.

**Global Interpretability:** Global insights into how features affect predictions across the entire dataset is required , not just for individual instances.

**Comparative Analysis:** A method that allows comparison of feature importance across different models or datasets reliably.

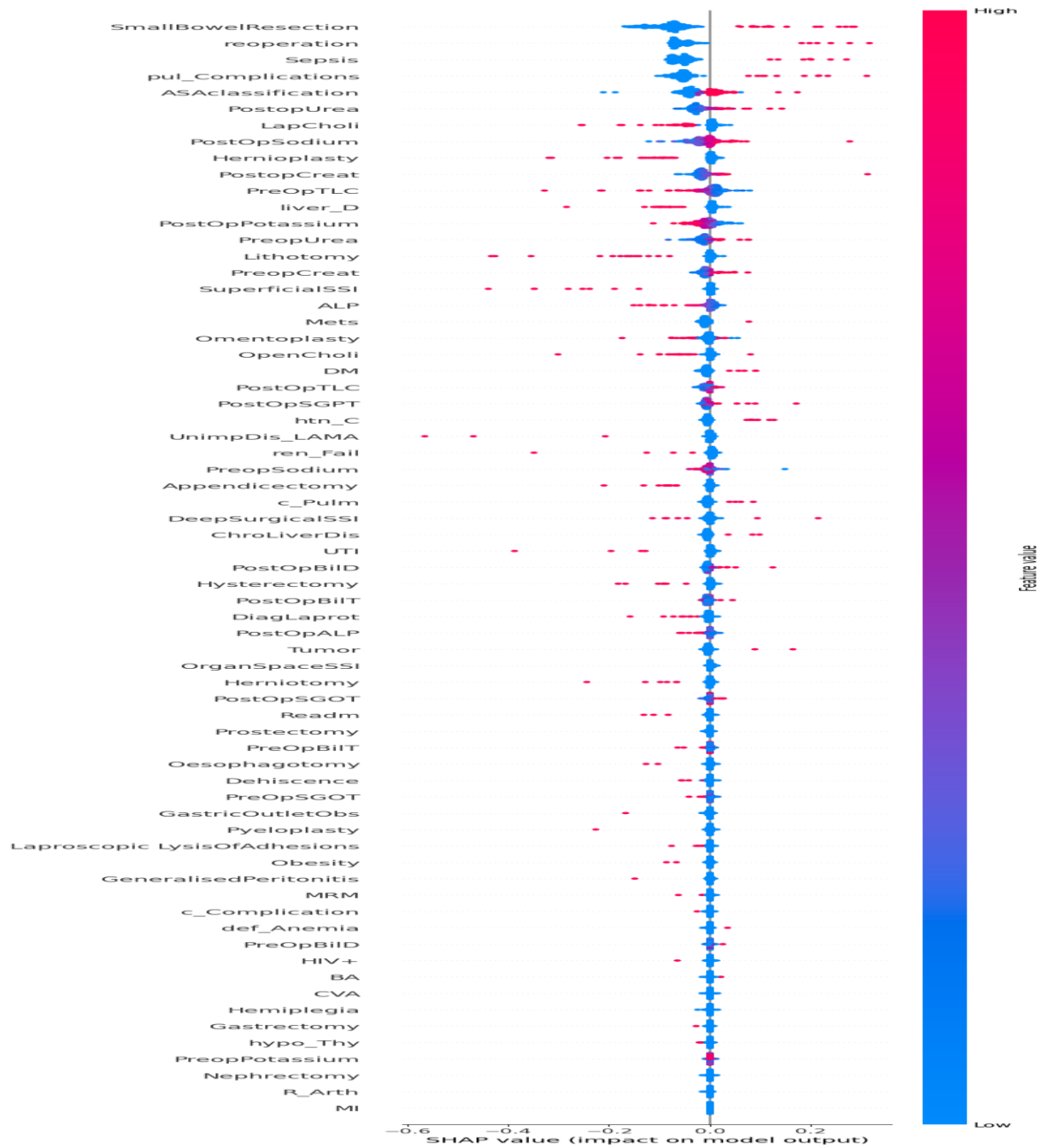
**Capturing Interaction Effects:** SHAP can capture complex interactions between features better than LIME, especially when using the TreeSHAP algorithm for tree-based models.

**Recommendation**

**For Global Interpretability:** SHAP is to be used , as it offers a comprehensive view of feature importance and their contributions.

**For Local Insights:** LIME is to be used for specific predictions where need is to explain individual cases.

In many cases, it can be beneficial to use both methods in tandem: SHAP for a broad understanding and LIME for deeper dives into specific predictions. This approach can provide a more rounded interpretation of your model's behavior.



**FIGURE 4.17: SHAP VALUES OF VARIABLES**

#### **4. Summary of Findings**

Prediction of mortality in surgical procedure is an important task for surgeon in doctor patient relationship. Various conventional techniques based on statistical methods are available including American Society of Anesthesiologists (American Society of Anaesthesiologists,1963) classification, Surgical Apgar Score, Surgical Risk Calculator (ACS-SRC) of American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP). Artificial Intelligence based models using Machine learning and Deep Learning were developed and compared using a small data set for their accuracy and reliability in predicting the outcome.

First objective of the study was to assess which method whether Machine Learning or Deep Learning is most capable of identifying mortality in a small dataset

The small dataset with sample size of 932 was highly skewed with serious class imbalance due to rare occurrence of mortality in hospital . Machine learning models were used with correction of class imbalance with variational auto encoder, and without correction of class imbalance (Original Data) but with the parameter `class_weight` "balanced" for certain models like Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. While the Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, and XGB Classifier show improved performance with variational Autoencoder transformed data, Logistic Regression and SVC did not perform well on the VAE data, indicating that simpler models may not be able to capture the complex patterns that the

VAE is generating. The Generative, Deep Neural Network (DNN) outperformed all machine learning all in terms of identifying True positives, achieving the best performance.

Second question was whether Generational autoencoder can be used using variational autoencoder to correct for class imbalance in training data with superior results. To understand that synthetic data with oversampling techniques like SMOTE and its derivatives were compared with variational Auto encoder augmented data in deep learning model for prediction of mortality

To evaluate the effectiveness of two oversampling methods—SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002) and its variants, along with VAE (Variational Autoencoder) (Kingma et al., 2014)—were employed to enhance the dataset for Deep Neural Network (DNN) models. The approaches utilized encompassed, Random Oversampling, SMOTE, BSSMOTE, Adasyn, Deep SMOTE, and VAE.

While Random Oversampling achieved high precision, it resulted in lower recall and a modest F1 score of 0.64. SMOTE improved recall(.77) but precision dropped to .56 with a F1 score of 0.65.BSMOTE(Borderline-SMOTE) showed F1 score of .71, ROC-AUC was also higher(.92) showing better distinguishing between classes. Adasyn resulted in a recall parameter of 0.77, meaning it missed many positives cases, precision dropped to 0.48, which indicated a higher number of false positives. Deep SMOTE achieved a strong performance across the board, with high recall (0.92) and precision (0.63). The ROC-AUC score of 0.97 was the highest, indicating excellent overall model performance in separating the classes. VAE-based model showed consistently high performance across multiple metrics. VAE transformed data produced a high F1 score of 0.71, a strong recall of 0.85, and a ROC-AUC score of 0.95. This showed that the model based on VAE-transformed data is effective at identifying the minority class while maintaining high accuracy and

precision The ROC-AUC of 0.95, though slightly lower than Deep SMOTE, was still very good but VAE's ability to model complex latent representations allows it to uncover hidden patterns in the data, made it most suitable. Further DNN model with Autoencoder for generation of synthetic data gave a probabilistic generative output deterministic in nature for each single run creating high uncertainty in predicting the outcome of mortality with no definite bounds of prediction due to two critical uncertainties: 1) epistemic uncertainty, accounting for the uncertainty in the model, and 2) aleatoric uncertainty, representing the impact of random disturbance, such as measurement errors. Accordingly, probabilistic models that measure the stochasticity and provide the range were tried. For Model development for prognostication of mortality, two approaches were tried. In the first approach, individual models were trained, which included a DNN model with VAE, a probabilistic generative model, two Probabilistic models with a Flipout layer at the end versus in all layers, and a Bayesian model. To generate the bounds the models were run 100 or more times and the results were aggregated. VAE model had strong performance across all metrics, with a balanced precision, recall, and F1 score of 0.77. The model achieved high accuracy (0.97), indicating that it correctly predicted most of the cases. The ROC-AUC score of 0.95 suggested excellent discriminatory ability between the classes. Of the probabilistic models, first model with Flipout at last layer underperformed compared to the VAE, with a significant drop in precision (.67), recall (.62), and F1 score (0.64). The accuracy was still relatively high (0.95), but the lower precision and recall suggested that it may struggle with classifying the positive class correctly. With ROC-AUC of 0.84, it was less capable of distinguishing between the classes compared to VAE. The Flipout in All Layers model performed poorly with extremely low accuracy (0.06) and precision (0.06), although its recall was 1.0. This meant that while it identified all the true positives, it did so at the expense of many false positives, resulting in poor overall performance. The

F1 score of 0.11 reflected this imbalance, with very low precision but perfect recall. The ROC-AUC of 0.30 suggested that the model did not differentiate between the two classes well, essentially performing no better than random guessing. The Bayesian model performed similarly to the Flipout Last Layer model, with an accuracy of 0.96 and higher precision (0.70), but lower recall (0.54). This indicated that the model was better at avoiding false positives than it was at capturing all true positives, which was desirable considering that the objective was to predict mortality in clinical settings. The F1 score of 0.61 reflected this trade-off between precision and recall, showing slightly lower performance compared to the Flipout Last Layer model. ROC-AUC of 0.80 is slightly lower than that of the Flipout Last Layer model (0.84), suggesting a moderate ability to differentiate between classes. Since probabilistic models did not perform well, and the aim was to move from deterministic approach to stochastic approach to better capture variability in data, explore a wider solution space, and provide probabilistic output, in the second approach Ensembles models were used where in combination of VAE, Model\_1, Model\_2, Bayesian were combined to develop Ensemble models which were then evaluated for accuracy, precision, recall, F1 score, True positive (TP), False positive (FP), False negative (FN), True positive (TP), Area Under the Curve, Mean probability where outcome is mortality i.e. 1, mean global probability, global entropy and mean entropy to select the best model for prediction of mortality in clinical settings. Ensemble model developed were

vae\_model + model\_1 + model\_2

vae\_model + model\_2 + bayesian\_model

vae\_model + model\_1 + bayesian\_model

bayesian\_model + model\_1 + model\_2

vae\_model + model\_1 + bayesian\_model + model\_2

vae\_model + model\_1

vae\_model + model\_2

bayesian\_model + model\_1

bayesian\_model + model\_2

model\_1 + model\_2

In numerous cases, conventional deep learning methods tend to generate overly confident probabilistic predictions, particularly when the training datasets are limited in size. This issue often becomes more pronounced, leading to complications when deep neural networks are used in contexts where accurate uncertainty quantification is essential. To address this, ensemble models were calibrated using temperature scaling, Platt scaling, and isotonic regression. Producing well-calibrated probabilistic predictions is vital for effective risk management, especially when decisions hinge on the outputs of these models. The effectiveness of the calibration techniques was evaluated using the Brier Score, which showed a notable decrease after calibration, indicating that the predicted probabilities became better aligned with the actual outcomes. Temperature scaling generally yielded strong performance across most ensemble models, maintaining high accuracy while achieving a good balance between precision and recall, making it a dependable method for model calibration. Conversely, Platt scaling encountered challenges with certain models, especially those incorporating Bayesian layers with Flipout, where it exhibited poor precision and F1 scores, although it performed adequately with simpler VAE-based ensembles.

Isotonic Regression came out as relatively stable and provided competitive performance close to Temperature Scaling. Post Calibration, the ensemble of VAE + Model\_1 (Flipout Last Layer) + Model\_2 (Flipout All Layers) stood out as the best performing model across all key metrics, calibration techniques, and uncertainty assessments for the following



reasons. Firstly, it had Balanced Metrics Across Precision(.65), Recall(.85), and F1 Score(.73),mean probability of positive class(.63+/-0.004).Post calibration the ensemble consistently achieved an F1 score, precision, and recall of 0.77, demonstrating a strong balance between precision (correctly identifying positive cases) and recall (capturing as many true positives as possible). Secondly, it had High Accuracy and ROC-AUC with a value of 0.94, Thirdly Low Entropy and Reduced Uncertainty ,since the ensemble exhibited low entropy values (Global- 5.4, Positive Class- 2.5), indicating that the model was making confident predictions. Fourthly it had Well-Calibrated Probabilities .The ensemble demonstrated strong performance in terms of calibration using all three techniques (Temperature Scaling, Isotonic Regression, Platt Scaling). Isotonic Regression yielded a Brier score of 0.0254, highlighting the ensemble's reliability in providing calibrated probabilities—important for real-world deployment where uncertainty in predictions needs to be managed effectively. Fifthly ,it is robust on Imbalanced Data. Given the imbalanced nature of the dataset, with a small number of positive cases, the ensemble showed robustness with low false positives (3) and false negatives (3). This indicated that it can handle skewed class distributions without sacrificing performance and lastly, it showed Consistency Across Various Metrics where unlike other ensembles that showed variability across calibration techniques or suffered from high uncertainty, this ensemble maintained consistent performance across metrics, making it more dependable for practical, high-stakes applications like forecasting mortality. This study emphasises the importance of Brier scores before and after calibration and underscores the importance of employing calibration techniques to improve model reliability in predicting probabilities. The effectiveness of temperature scaling and isotonic regression is evident, with both techniques providing valuable adjustments to probability estimates. Overall, this analysis highlights the need for careful calibration to enhance the performance of ensemble models,

especially in contexts where prediction accuracy is critical. These insights not only validate the calibration techniques employed but also contribute to the ongoing discussion on improving model performance in machine learning, particularly in challenging environments characterized by class imbalance

The third question was whether the variable of importance derived from model explainability techniques of Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations, commonly known as SHAP can explain and interpret the model for prognostication of postoperative mortality with identical results?

There was considerable amount of match between variables of importance in two methods of interpretability. Both LIME and SHAP identified similar features impacting the mortality in surgical procedures. Some of the features negatively impacting the outcome were shared by both for example obesity  $<0$  with coefficient 1.1906 in LIME and  $-0.0005436729122828631$  in SHAP, Alkaline Phosphate  $0.11$  to  $\leq 0.14$  with coefficient  $-0.002194086349926584$  and  $-0.0137$  in SHAP. Omentoplasty feature with value  $\leq 0$  has a LIME coefficient of  $-0.45$  whereas SHAP has a weight of  $-0.011$ , Postop Direct Bilirubin with value with  $>.04$  has a LIME coefficient of  $-0.11$  and a SHAP weight of  $.009$ , PreopTLC  $<.11$  had a LIME coefficient of  $.20$  whereas PreopTLC  $>.19$  has a negative LIME coefficient of  $.17$  whereas SHAP weight for the feature is  $-.006$ . The same is true for other variables like Postop Bilirubin Total (LIME if  $>.08$ , a coefficient of  $-.06$  and SHAP weight of  $-.006$ , Diagnostic Laprotomy  $>0$ , a LIME weight of  $-0.11$  and SHAP weight of  $-.004$ , PreopSodium  $.62 \leq .68$ , a LIME Coefficient of  $.011$  and SHAP weight of  $-.004$ , PreopUrea  $\leq .07$ , a LIME coefficient of  $-0.015$  and a SHAP weight of  $-.004$ . Lap Choli  $\leq 0$  has a LIME coefficient of  $.93$  whereas SHAP has a weight of  $-.004$  meaning the same thing. For patients who did not undergo Laparoscopic Cholecystectomy, the LIME model

assigns a coefficient of 0.97. This suggests that not having the surgery is associated with a higher risk of the outcome (mortality) whereas SHAP weight is -0.047. For feature Chronic Pulmonary disease <0, LIME coefficient is .75 whereas SHAP value is -0.02.

In this study, both LIME and SHAP were applied to understand the factors impacting surgical mortality. Despite differences in how variable weights are calculated, both methods identified several common features influencing mortality, such as Sepsis, Postoperative Urea, Small Bowel Resection, ASA Classification, Postoperative SGPT, Chronic Liver Disease, Preoperative Creatinine, Postoperative Creatinine, Tumor: Reoperation, obesity, alkaline phosphatase levels, omentoplasty, and postoperative bilirubin etc. These features were consistently highlighted across both methods, reinforcing their significance in predicting patient outcomes. However, LIME provided more local insights, focusing on individual predictions, while SHAP offered a global perspective across all instances. For example, LIME emphasized sepsis as a crucial factor with a high positive impact, while SHAP showed a more balanced view with lower average importance. Some discrepancies, such as reoperation and tumor, were also observed, with LIME indicating a positive contribution while SHAP suggested a negative impact.

The analysis revealed that using LIME is beneficial for specific case-by-case interpretability, while SHAP excels in offering a comprehensive overview of feature importance across the entire dataset. Both methods together provided a thorough understanding of how various features influence surgical mortality, and their combined use is recommended for more accurate model interpretation.

## 4.5 Conclusion

Deep Learning model outperformed traditional Machine Learning techniques. By integrating insights from all three models -Variational Auto Encoder, Probabilistic models with flipout and the Bayesian Model , it became possible to leverage the strengths of each model to enhance overall predictive performance and reliability.

VAE Model: Demonstrated robust overall performance and provided a solid baseline for mortality predictions.

Probabilistic Model with Flipouts: Offered valuable insights into prediction uncertainty, which is critical for assessing the reliability of predictions in clinical settings.

Bayesian Model: Highlighted areas of improvement in precision and contributed to refining the model's interpretability

Together, these models enhanced the reliability and accuracy of mortality predictions, leading to the development of an ensemble model incorporating the VAE, Model\_1, and Model\_2 as the best predictor of mortality. This study emphasises the importance of Brier scores before and after calibration and underscores the importance of employing calibration techniques to improve model reliability in predicting probabilities. The effectiveness of temperature scaling and isotonic regression is evident, with both techniques providing valuable adjustments to probability estimates. Overall, this analysis highlights the need for careful calibration to enhance the performance of ensemble models, especially in contexts where prediction accuracy is critical. These insights not only validate the calibration techniques employed but also contribute to the ongoing discussion on improving model

performance in machine learning, particularly in challenging environments characterized by class imbalance. With respect explainability of models, the two methods used LIME and SHAP identified several common features influencing mortality, such Sepsis, Postoperative Urea, Small Bowel Resection, ASA Classification ,Postoperative SGPT, Chronic Liver Disease, Preoperative Creatinine, Postoperative Creatinine, Tumor: Reoperation , obesity, alkaline phosphatase levels, omentoplasty, and postoperative bilirubin etc. These features were consistently highlighted across both methods, reinforcing their significance in predicting patient outcomes. However, LIME provided more local insights, focusing on individual predictions, while SHAP offered a global perspective across all instances. For example, LIME emphasized sepsis as a crucial factor with a high positive impact.

## CHAPTER V:

### Discussion of Results

#### 5.1 Discussion of Results

The prediction of mortality in surgical procedures represents a crucial intersection of clinical decision-making and patient safety. This study aimed to evaluate the effectiveness of various machine learning and deep learning models in predicting surgical mortality using a small dataset characterized by significant class imbalance. By employing traditional statistical methods alongside advanced AI techniques, the research provides insights into the capabilities and limitations of these models in a healthcare context.

**Machine Learning Versus Deep Learning:** The first objective of the study was to determine which methodology—machine learning or deep learning—was more effective in identifying mortality outcomes. Despite the small sample size of 932, machine learning models that employed techniques to address class imbalance i.e. Variational Autoencoders (VAE), showed promising results. The Decision Tree Classifier, Random Forest Classifier, and Gradient Boosting Classifier performed better with the VAE-transformed data. However, simpler models like Logistic Regression struggled to capture the complex patterns generated by the VAE, suggesting that a more sophisticated approach is necessary for datasets with significant class imbalance.

The deep learning model, particularly the Generative Deep Neural Network (DNN), outperformed the machine learning models, achieving the highest true positive identification rates. This underscores the potential of deep learning techniques to extract intricate features and relationships in data, which may be particularly beneficial in clinical settings where accurate risk stratification is essential.

**Impact of Synthetic Data and Oversampling Techniques:** The second objective explored the efficacy of various oversampling methods, including SMOTE and its derivatives, compared to VAE-generated synthetic data. While traditional methods like Random Oversampling showed high precision, they resulted in lower recall. In contrast, Deep SMOTE demonstrated a strong overall performance with high recall and precision. The VAE, however, maintained a high F1 score, indicating its robustness in identifying the minority class while achieving good precision. This aligns with the study's findings that VAE-based models can effectively manage class imbalance without compromising the accuracy of predictions.

The exploration of epistemic and aleatoric uncertainty through probabilistic models further enriched the analysis. The probabilistic outputs from the ensemble models highlighted the inherent uncertainties in predicting surgical mortality, revealing the necessity of reliable risk assessment tools in clinical environments.

**Ensemble Modeling and Calibration Techniques:** The investigation into ensemble models revealed that combining VAE, Bayesian models, and other deep learning models resulted in improved performance metrics across the board. The calibrated ensemble model with the best balance of precision and recall underscores the importance of integrating

multiple predictive strategies. The successful application of calibration techniques, particularly temperature scaling and isotonic regression, illustrates their significance in refining model predictions and enhancing reliability. By reducing the Brier score post-calibration, the study underscores the need for accurate probability estimates in clinical decision-making.

**Interpretability of Models:** Lastly, the study's examination of model interpretability through LIME and SHAP demonstrated that both techniques yielded similar insights regarding important variables influencing surgical mortality. The alignment of variable importance between these methods enhances the credibility of the findings, providing clinicians with actionable insights into the factors contributing to mortality risk.

In summary, this research highlights the potential of advanced machine learning and deep learning techniques to improve the prediction of mortality in surgical procedures. By addressing class imbalance, employing effective calibration techniques, and utilizing interpretable models, the study contributes valuable knowledge to the field of surgical risk assessment, ultimately aiming to enhance patient outcomes through informed clinical decision-making.

## **5.2 Discussion of Research Question One**

Deep learning techniques outperformed traditional machine learning methods, including logistic regression, K-nearest neighbors (KNN), decision trees, random forests, gradient boosting, and XGBoost, in predicting mortality when utilizing Variational Autoencoders with small datasets. In a study by Wang et al. (2020), a model was developed for the early



diagnosis of Parkinson's disease (PD) using a limited dataset comprising 183 healthy individuals and 401 patients in the early stages of PD. The authors employed both deep learning and various other machine learning techniques to differentiate between Parkinson's patients and healthy controls. They noted that methods such as logistic regression (LOGIS), penalized logistic regression (LOGIS\_PEN), random forest (RF), discriminant analysis (DIS), KNN, support vector machines (SVM), and classification trees were efficient. However, the deep learning model demonstrated superior detection capabilities, achieving an impressive accuracy of 96.45%. This high performance was largely attributed to the deep learning model's ability to automatically learn both linear and nonlinear features from PD data without requiring manual feature extraction.

In this study, variational autoencoder was used to correct the serious class imbalance. The study supports the facts that VAE transformed data when applied to traditional machine learning and deep learning showed better performance by DNN model, particularly in improving recall by way of comparison. The Generative, Deep Neural Network (DNN) outperformed them all in terms of identifying True positives, achieving the best performance. The closely running model was XGB classifier which had an accuracy of .96, precision of .71 and a recall of .71 with F1 score of .66, as against accuracy of .96, precision of .61, recall of .85 and F1 score of .71, whose performance was not equal in terms of identifying True positives a major concern in predicting mortality.

## 5.2 Discussion of Research Question Two

Whether Generational autoencoder can be used using variational autoencoder to correct for imbalance in training data with superior results?

The Variational Autoencoder (VAE) achieved the best overall performance with high precision and recall, making it a strong candidate for cases requiring a balance between detecting positives accurately and minimizing false positives and negatives. Fajardo et al. (2018) described a method extending Variational Autoencoders (VAEs) to address imbalanced data. which showed that the new method outperformed SMOTE with respect to a downstream binary classification task. The study compared results across three classifiers—logistic regression (LR), random forest (RF), and multi-layer perceptron (MLP) and found that oversampling with VAE led to improved accuracy metrics on the test set.). Oversampling with VAE significantly outperformed SMOTE and helped the classifier to achieve outstanding accuracy metrics on the test set. In particular, when using LR and MLP, the precision and F1 scores of the VAE were significantly higher. In their 2021 study, Islam et al. observed that the Variational Autoencoder (VAE) outperformed other data augmentation techniques. When compared to SMOTE, VAE led to an 8% and 4% improvement in specificity for VAE-Logistic Regression (LR) and VAE-Support Vector Machine (SVM) models, respectively. Similarly, when compared to ADASYN, the sensitivity increased by 6% for VAE-LR and 5% for VAE-SVM, highlighting the effectiveness of VAE in enhancing model performance.

In the current experiment, VAE demonstrated superior performance, outperforming other methods both in data augmentation and mortality prediction by accurately identifying

the target. VAE also made the most significant contribution in ensemble models. However, while VAE can provide an overall measure of uncertainty, it cannot quantify uncertainty at the individual case level. This limitation highlights the need to combine VAE with probabilistic models to obtain a range of probabilities for each outcome. By integrating insights from both models, it became possible to achieve a more comprehensive understanding of the VAE model's performance and the uncertainty in its predictions.

Amodei et al. (2016) emphasized that although neural networks (NNs) can achieve high accuracy in supervised learning tasks, they often fail to effectively quantify predictive uncertainty, leading to overconfident predictions. Such overconfidence in incorrect predictions can have serious or even harmful consequences, underscoring the importance of proper uncertainty quantification in practical applications. The paper identifies AI safety concerns, particularly regarding overconfident neural network outputs, and stresses the need for uncertainty quantification in real-world scenarios.

In healthcare, VAEs have been shown to successfully model complex data patterns, as seen in medical image analysis (Gondara, 2016), making them a promising tool for mortality prediction in complex datasets. "Generative models like VAEs offer a novel solution for imbalanced datasets in healthcare by capturing latent data distributions, providing an alternative to traditional methods like Deep SMOTE (Xu & Goodacre, 2018)." **Flipout Layers Only at the End** has the advantages that it Captures uncertainty mainly in the final decision-making layer, which can be sufficient for many applications, which have **reduced Complexity** and thereby is simpler to implement compared to adding Flipout layers in all layers and lastly has **Lower Computational Cost** being Less

computationally expensive than applying Flipout layers throughout the network. The disadvantage is that it is **less Comprehensive and hence** does not capture all sources of uncertainty as effectively as applying Flipout layers throughout the network (Wen, et al., 2018). Bayesian Neural Networks (BNNs) estimate weights as probabilistic distributions, enabling them to account for uncertainty in predictions (Joshi and Dhar, 2022). According to Blundell et al. (2015), weights with higher uncertainty introduce greater variability into the network's decisions, naturally encouraging exploration. As the network observes more data, this uncertainty decreases, allowing the decisions to become more deterministic as the model gains a better understanding of the environment.

In his study on uncertainty quantification and deep ensembles, Rahaman (2021) addressed the challenge of calibrating deep ensembles and explored the interaction between three commonly used methods for adapting deep learning to low-data scenarios: ensembling, temperature scaling, and mixup data augmentation. Using standard neural architectures, such as ResNet18 for CIFAR10/100 and ResNet34 for ImageNet, Imagewoof, and Diabetic Retinopathy datasets, Rahaman found that combining these models into an ensemble improved calibration, resulting in better-calibrated predictions across different tasks.

BNN estimates weights in the form of probabilistic distributions and thereby could account for uncertainty in the predictions (Joshi, and Dhar, 2022). As per Blundel et al. (2015), weights with greater uncertainty introduce more variability into the decisions made by the network, leading naturally to exploration. As more data are observed, the uncertainty can decrease, allowing the decisions made by the network to become more deterministic

as the environment is better understood. Rahaman (2021) in his paper on Uncertainty quantification and deep ensembles. *Advances in neural information processing systems*, discussed the problem of calibrating deep-ensembles and examined the interaction between three of the most simple and widely used methods for adopting deep-learning to the low-data regime: ensembling, temperature scaling, and mixup data augmentation. He used standard neural architectures networks , For CIFAR10/100 , it was ResNet18 and for ImageNet/Image woof , and Diabetic Retinopathy it was ResNet34. When these models were pooled to make an ensemble, the pooled model was better calibrated. Finally, pool-then-calibrate with temperature scaling had the best performance (Niculescu-Mizil & Caruana, 2005).Further, calibration techniques like isotonic regression and Platt scaling may perform poorly on imbalanced data. It mentions that isotonic regression can overfit when applied to small or imbalanced datasets due to its flexibility. This study also found that pooled ensembled model followed by calibration did better in terms of identification of positives particularly VAE model plus Model\_1 and Model\_2. **Guo et al., (2017)** in their study "**On Calibration of Modern Neural Networks**" noted that typical calibration techniques tend to fail with imbalanced data, especially when there are very few positive examples, and calibration can even degrade the classifier's performance for minority classes. Further, Calibration techniques like isotonic regression are prone to overfitting, especially when the data is limited or imbalanced, as they require a large amount of data to create reliable mappings between predicted and true probabilities. This study provides insightful findings that both align with and challenge existing literature on calibration techniques for imbalanced datasets. Pool-then-calibrate with temperature scaling had the

best performance, aligning well with Niculescu-Mizil & Caruana's (2005) discussion on the importance of model calibration in improving probability predictions. This emphasizes the effectiveness of a systematic approach to calibration after ensembling, particularly in scenarios where class imbalance is a concern. It suggests that temperature scaling not only enhances the predictions of ensemble models but also addresses the issues commonly faced with minority class predictions. However contrary to the Guo et al.(2017) observation ,brier score after isotonic regression were smaller was not confirmed as temperature scaling as also Isotonic regression reduced the brier score for all models even though values of brier scores were moderate for temp scaling and quite significant for isotonic regression. Both **temperature scaling and isotonic regression effectively mitigated calibration issues for the minority class**, which supports the idea that careful application of these techniques can yield more reliable results even in challenging data conditions. This adds to the discourse on calibration by indicating that with the right methods, performance degradation for minority classes can be addressed. **Brier scores were moderate for temperature scaling but quite significant for isotonic regression** further underscoring the complexities in using these calibration techniques. It suggests that while temperature scaling may provide a robust performance across models, isotonic regression may offer superior calibration in specific cases. The distinction in performance levels also emphasizes the importance of selecting calibration techniques based on the specific characteristics of the dataset and the models used. These insights not only validate the calibration techniques but also contribute to the ongoing discussion on improving model

performance in machine learning, particularly in challenging environments characterized by class imbalance.

### 5.3. Discussion of Research Question Three

**RQ3.** Whether the variable of importance derived from model explainability techniques of Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations, commonly known as SHAP can explain and interpret the model for prognostication of postoperative mortality with identical results?

There was considerable amount of match between variables of importance in two methods of interpretability. Both LIME and SHAP identified similar features impacting the mortality in surgical procedures. Some of the features negatively impacting the outcome were shared by both for example obesity  $<0$  with coefficient 1.1906 in LIME and  $-0.0005436729122828631$  in SHAP, Alkaline Phosphate  $0.11$  to  $\leq 0.14$  with coefficient  $-0.002194086349926584$  and  $-0.0137$  in SHAP. Omentoplasty feature with value  $\leq 0$  has a LIME coefficient of  $-0.45$  whereas SHAP has a weight of  $-0.011$ , Postop Direct Bilirubin with value with  $>.04$  has a LIME coefficient of  $-0.11$  and a SHAP weight of  $.009$ , PreopTLC  $<.11$  had a LIME coefficient of  $.20$  whereas PreopTLC  $>.19$  has a negative LIME coefficient of  $.17$  whereas SHAP weight for the feature is  $-.006$ . The same is true for other variables like Postop Bilirubin Total (LIME if  $>.08$ , a coefficient of  $-.06$  and SHAP weight of  $-.006$ , Diagnostic Laprotomy  $>0$ , a LIME weight of  $-0.11$  and SHAP weight of  $-.004$ , PreopSodium  $.62 \leq .68$ , a LIME Coefficient of  $.011$  and SHAP weight of  $-.004$ , PreopUrea  $\leq .07$ , a LIME coefficient of  $-.015$  and a SHAP weight of  $-.004$ . Lap Choli  $\leq 0$  has a LIME coefficient of  $.93$  whereas SHAP has a weight of  $-.004$  meaning the same

thing . For patients who did not undergo Laparoscopic Cholecystectomy, the LIME model assigns a coefficient of 0.97. This suggests that not having the surgery is associated with a higher risk of the outcome (mortality) whereas SHAP weight is -0.047. For feature Chronic Pulmonary disease <0 ,LIME coefficient is .75 whereas SHAP value is -0.02.

There are however a few differences. Reoperation >0 had a Lime coefficient of .25 whereas SHAP had a negative weight of -0.008. In case of Tumor >0 , LIME assigned a value of .30 whereas SHAP had a negative weight of -0.003. In this study, both LIME and SHAP were applied to assess feature importance in predicting postoperative mortality. Despite differences in how the two methods calculate variable weights, there was considerable overlap in the features identified as most important by both techniques. Features such as **obesity**, **alkaline phosphatase levels**, **omentoplasty**, and **postoperative bilirubin** were highlighted by both LIME and SHAP as having significant impacts on mortality outcomes.

For instance, obesity had a LIME coefficient of 1.1906 and a SHAP weight of -0.0005, both suggesting it negatively impacts patient outcomes. Similar alignment was found for **alkaline phosphatase** (LIME coefficient: -0.0021, SHAP: -0.0137) and **omentoplasty** (LIME: -0.45, SHAP: -0.011). These findings indicate that both methods can identify crucial variables despite using different interpretative approaches.

However, some discrepancies were observed between the two methods. For example, **reoperation** had a LIME coefficient of 0.25, indicating a positive impact, while SHAP assigned a negative weight of -0.008. Similarly, for **tumor** presence, LIME suggested a positive contribution (0.30), whereas SHAP indicated a negative effect (-0.003). These



differences reflect the distinct ways in which LIME and SHAP calculate feature importance—LIME providing local interpretability for individual predictions and SHAP offering a global view across all instances.

On the whole, both methods consistently identified key features like **sepsis**, **postoperative urea**, **small bowel resection**, **ASA classification**, and **chronic liver disease** as significant contributors to surgical mortality outcomes. LIME tends to emphasize individual predictions, offering more localized insights, while SHAP provides a broader, more balanced perspective on feature importance across the entire dataset. For example, LIME assigned a higher coefficient to **sepsis** (2.32), signalling its strong local impact, while SHAP presented a more moderate contribution (0.09), reflecting its average importance across all predictions.

This study's findings align with previous research by Bandstra et al. (2023), who observed comparable results between LIME and Kernel SHAP. They noted that SHAP is a generalization of LIME, which explains the similarity in their results. In conclusion, while LIME and SHAP approach interpretability differently, both methods provide valuable insights into the factors affecting postoperative mortality. Using both methods together offers a comprehensive understanding, with LIME aiding in case-specific analysis and SHAP providing a global assessment of feature importance. Omentoplasty >0 .087 and SHAP had a feature weight of -.011.

While there was a dissimilarity in how the variables weights are calculated in LIME and SHAP ,there is considerable amount of match between variables of importance in two methods of interpretability. Both LIME and SHAP has identified similar features impacting

the mortality in surgical procedures .Some of the features negatively impacting the outcome are shared by both. There are however a few differences .Reoperation >0 has a Lime coefficient of .25 whereas SHAP has a negative weight of -.008. In case of Tumor >0 ,LIME assigns a value of .30 whereas SHAP has a negative weight of -.003. Omentoplasty >0 .087 and SHAP has a feature weight of -.011.

On the whole while both LIME and SHAP had differences in how variable weights were calculated, both methods identified several common features influencing mortality, These features were consistently highlighted across both methods, reinforcing their significance in predicting patient outcomes. , such as **Sepsis, Postoperative Urea, Small Bowel Resection: ASA Classification ,Postoperative SGPT, Chronic Liver Disease, Preoperative Creatinine, Postoperative Creatinine, Tumor, Reoperation** , obesity, alkaline phosphatase levels, omentoplasty, postoperative bilirubin etc. These features were consistently highlighted across both methods, reinforcing their significance in predicting patient outcomes. However, LIME provided more local insights, focusing on individual predictions, while SHAP offered a global perspective across all instances. For example, LIME emphasized sepsis as a crucial factor with a high positive impact, while SHAP showed a more balanced view with lower average importance. Some discrepancies, such as reoperation and tumor, were also observed, with LIME indicating a positive contribution while SHAP suggested a negative impact. The analysis revealed that using LIME is beneficial for specific case-by-case interpretability, while SHAP excels in offering a comprehensive overview of feature importance across the entire dataset. Both methods together provided a thorough understanding of how various features influence surgical

mortality, .Bandstra et al.(2023) in their study found that LIME and Kernel SHAP gave comparable and nearly identical results. They noted that this coincidence is not surprising given that SHAP is a generalization of LIME.This study replicates the observation.

## CHAPTER VI:

### **SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS**

#### **6.1 Summary**

This study demonstrates that data transformation using a Variational Autoencoder (VAE) yields superior results for predictive models when utilizing complex machine learning and deep neural networks. In comparison, simpler machine learning models fail to capture the intricate patterns produced by the VAE. The VAE-based models consistently outperformed other techniques, including SMOTE and Deep SMOTE, across multiple metrics. While both the Deep SMOTE and VAE models showed similar performance in terms of F1 score and recall, the VAE had several distinct advantages that made it a better choice for mortality prediction. These advantages include its ability to model complex latent representations, which enables it to uncover hidden patterns and adapt to various datasets. Additionally, the generative capabilities of the VAE offer a novel approach to addressing class imbalance and uncertainty, providing greater flexibility for research and real-world applications. Unlike Deep SMOTE, which directly augments the dataset, the VAE learns a latent representation of the data distribution, allowing it to handle imbalance through complex generative modeling. To handle the uncertainty and variability in predictions, ensemble models were developed, combining different models (VAE, Flipout, and Bayesian). These ensembles were further calibrated using techniques like Temperature Scaling, Platt's Scaling, and Isotonic Regression to improve probabilistic output reliability.

Calibration reduced Brier Scores, demonstrating better alignment of predicted probabilities with actual outcomes.

The final ensemble of VAE + Model\_1 (Flipout Last Layer) + Model\_2 (Flipout All Layers) emerged as the best-performing model across all key metrics. It achieved high accuracy (0.94), a balanced F1 score (0.73), and low uncertainty, making it a robust choice for mortality prediction. The ensemble's strong calibration and reliability make it suitable for real-world clinical applications, providing well-calibrated and dependable probabilistic predictions.

Both **temperature scaling and isotonic regression effectively mitigated calibration issues for the minority class**, which supports the idea that careful application of these techniques can yield more reliable results even in challenging data conditions. This adds to the discourse on calibration by indicating that with the right methods, performance degradation for minority classes can be addressed. This study found a high degree of overlap between the variables identified as important by two interpretability methods: LIME and SHAP. Both methods highlighted similar features impacting mortality in surgical procedures, such as obesity, Alkaline Phosphate levels, and certain pre- and postoperative lab values. For instance, obesity was shown to negatively impact outcomes with similar coefficients in both LIME and SHAP.

Although most feature weights were aligned, some discrepancies were observed. For example, LIME assigned a positive coefficient for reoperation and tumor presence, while SHAP indicated negative weights for the same features. In these cases, LIME's values appeared more appropriate. The findings align with previous research by Bandstra et al. (2023), which reported comparable results between LIME and Kernel SHAP. This suggests that despite differences in calculation methods, approaches can provide consistent insights into feature importance in clinical settings

## **6.2 Implications**

The implications of this study on the prediction of mortality in surgical procedures are significant and multifaceted, particularly within the context of enhancing clinical decision-making and patient management. Here are several key points to consider:

### **6.2.1. Enhanced Predictive Accuracy**

The study demonstrates that advanced machine learning and deep learning techniques, particularly those utilizing Variational Autoencoders (VAE) and ensemble methods, can significantly improve the accuracy of mortality predictions in surgical settings. By leveraging complex models that can learn from imbalanced datasets, healthcare providers can make more informed decisions about patient risk, leading to better outcomes.

### **6.2.2. Clinical Decision Support**

The development of robust predictive models offers valuable support for surgeons and clinical teams. By integrating these models into clinical workflows, healthcare professionals can better assess the risks associated with surgical procedures, ultimately leading to tailored preoperative assessments and interventions. This could enhance discussions with patients regarding their treatment options and expectations.

### **6.2.3. Addressing Class Imbalance**

The study highlights the importance of addressing class imbalance in mortality prediction. Traditional models often struggle with this issue, leading to poor performance in identifying rare events like mortality. By employing techniques like VAE and various oversampling methods, the study presents a pathway to improve model reliability and accuracy, paving the way for more equitable predictive capabilities across diverse patient populations.

#### **6.2.4. Calibration Techniques for Reliable Probabilities**

The emphasis on calibration techniques, such as temperature scaling and isotonic regression, underscores the necessity of producing well-calibrated probabilistic predictions in clinical settings. Accurate probability estimates are crucial for effective risk management and can enhance the trust that clinicians and patients place in model predictions, thereby improving the overall utility of predictive analytics in healthcare.

#### **6.2.5. Interpretable and Explainable Models**

The alignment of variable importance between Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) reinforces the value of explainability in predictive modelling. Understanding which features influence mortality risk helps clinicians interpret model outputs and fosters confidence in the decision-making process. This interpretability is essential for translating model predictions into actionable clinical strategies.

#### **6.2.6. Integration into Surgical Practice**

The findings encourage the integration of machine learning and deep learning models into routine surgical practice. By embracing these innovative approaches, hospitals and surgical canters can enhance their predictive capabilities, improve patient stratification, and allocate resources more effectively, ultimately leading to improved surgical outcomes and reduced mortality rates.

#### **6.2.7. Future Research Directions**

This study opens avenues for future research to explore additional variables, enhance model architectures, and test these predictive models in larger, more diverse cohorts. There

is potential for further refinement of algorithms to address limitations identified in current models, as well as opportunities to explore real-time data integration and adaptive learning approaches in surgical settings.

#### **6.2.8. Broader Implications for Healthcare Analytics**

Beyond surgical mortality prediction, the methodologies and insights gained from this study have broader implications for healthcare analytics. The techniques developed can be applied to other clinical scenarios, potentially improving outcomes in areas like critical care, oncology, and other high-stakes medical domains where accurate risk assessment is paramount.

Overall, the implications of this study underscore the transformative potential of machine learning and deep learning in improving predictive analytics within healthcare. By focusing on enhancing accuracy, reliability, interpretability, and integration into clinical practice, these advancements can significantly contribute to better patient care and outcomes in surgical settings.

### **6.3 Recommendations for Future Research**

Future research should focus on implementing and evaluating calibration error metrics such as Expected Calibration Error (ECE) and Maximum Calibration Error (MCE). These metrics will help identify where the model predictions deviate from true probabilities, offering deeper insights into the reliability of probabilistic outputs across different prediction intervals. The findings regarding the effectiveness of temperature scaling and the unexpected performance of isotonic regression challenge some established assumptions, indicating that these techniques can be beneficial even in scenarios

traditionally viewed as problematic. This encourages further exploration of calibration methods and their adaptability across different datasets and modeling scenarios. Additionally, increasing data size should be a priority. This can be achieved by obtaining and incorporating more samples or leveraging external data sources. Expanding the dataset would reduce overfitting and improve model generalization, which is particularly crucial for VAE-based models. These models often perform better with larger datasets due to their high capacity for learning complex patterns.

Furthermore, applying data augmentation techniques, such as using other generative models like Generative Adversarial Networks (GANs) or time-series augmentation methods for sequential data, can synthetically increase data size and diversity. This approach can help enhance model robustness and provide a richer data landscape for improved predictive performance.

#### **6.4 Conclusion**

A small dataset of 930 patients was used to prognosticate the mortality. A Deep neural network model with synthetic data from Variational Autoencoder was found to be superior to conventional Machine Learning techniques including Logistic regression, K-Neighborhood Classifier, SVC, DecisionTreeClassifier, Random Forest Classifier, Gradient Boosting classifier and XGBCLASSIFIER with better recall, precision,Fi score and ROCAUC Values. Since deterministic models may not replicate the results due to Aleatoric and Epistemic uncertainty,03 probabilistic Probabilistic models with flipout at the last layer, in all layers and Bayesian model was added to measure uncertainty. Since



these models except for probabilistic model with flipout did not develop desired statistics due to lack of real data, following Ensemble models were developed:

1. vae\_model + model\_1 + model\_2
2. vae\_model + model\_2 + bayesian\_model
3. vae\_model + model\_1 + bayesian\_model
4. bayesian\_model + model\_1 + model\_2
5. vae\_model + model\_1 + bayesian\_model+model\_2
6. vae\_model + model\_1
7. vae\_model + model\_2
8. bayesian\_model + model\_1
9. bayesian\_model + model\_2
10. model\_1 + model\_2

These ensembles were further calibrated using techniques like Temperature Scaling, Platt's Scaling, and Isotonic Regression to improve probabilistic output reliability. Calibration reduced Brier Scores, demonstrating better alignment of predicted probabilities with actual outcomes.

Post calibration, VAE vae\_model + model\_1 + model\_2 was found to be best performer. As these Black box models are not interpretable, two techniques of model interpretability i.e. LIME and SHAP were used. Both gave almost similar results with variables of Sepsis, PostOPUrea, Small Bowel Resection, ASA Classification, Post op SGPT, Chronic Liver

disease ,Preop Creatinine, Postop creatinine, Tumor, Reoperation ,Postop Sodium, Mets ,Omentoplasty, DM ,Pulmonary complication, Postop Bilirubin Direct and LapCholi being significant influencers of the model for prognostication of mortality. While the variables were similar, their order in terms of hierarchy of influence was different .Similarly coefficients of LIME and SHAP were different due to different methodology adopted in their calculation. The first 10 variables in descending order as per the LIME are Sepsis > 0.00: weight .36, Postop Urea > >0.13: 1.10, Small Bowel Resection > 0.00: 1.05, ASA classification .50 to <= 1.00: 0.50,Postop SGPT > 0.02: .37, Chro Liver Dis > 0.00: 0.33, Pre op Creat > 0.06: 0.32 , Post op Creat > 0.10: 0.31, Tumor > 0.00: 0.30, reoperation > 0.00: 0.25, Post Op Sodium > 0.67: 0.21, PreopUrea > 0.14: 0.19 ,Mets > 0.00: 0.16, Post Op SGOT > 0.05: 0.11 ,Omentoplasty > 0.00: 0.08 , DM > 0.00: 0.03 and Pul Complications > 0.00: 0.03 , 0.02 < PostOpBilD <= 0.04: 0.01, LapCholi <= 0.00: 0.97. Whereas same 10 variables of importance has mean SHAP values for Sepsis : .09 , PostopUrea: 0.009,Small Bowel Resection: 0.18, ASA classification: 0.02, Post Op SGPT: 0.04, ChroLiverDis: 0.03,PreopCreat: .006,PostopCreat:.002, Tumor:-.003, reoperation : -0.008 , Post Op Sodium: .011,PreopUrea .009 , ,Mets: .0024,PostopSGOT:.004,Omentoplasty:.011, DM: 0.017, pul\_Complications: 0.02. Post Op BilD: 0.007, LapCholi, -0.011. Both **temperature scaling and isotonic regression effectively mitigated calibration issues for the minority class**, which supports the idea that careful application of these techniques can yield more reliable results even in challenging data conditions. This adds to the discourse on calibration by indicating that with the right methods, performance degradation for minority classes can be addressed.

This encourages further exploration of calibration methods and their adaptability across different datasets and modelling scenarios. Future research should emphasize the implementation and evaluation of calibration error metrics such as Expected Calibration Error (ECE) and Maximum Calibration Error (MCE). These metrics are crucial for pinpointing where model predictions diverge from true probabilities, providing valuable insights into the trustworthiness of probabilistic outputs across various prediction intervals. In addition to refining calibration metrics, expanding the dataset is essential for improving model performance. This can be accomplished by acquiring more samples or incorporating external data sources, both of which are critical for deep learning models, which are known to perform best with larger datasets.

Moreover, applying advanced data augmentation techniques should be a priority. Incorporating methods such as Generative Adversarial Networks (GANs) alongside traditional approaches can generate synthetic data that closely mimics real-world distributions, enhancing the model's ability to generalize. These generative models not only increase the dataset size but also enrich the diversity of the data, helping deep learning models to learn better representations. This combination of robust calibration methods, increased data, and sophisticated augmentation strategies will lead to more reliable predictions and superior performance across all models.

## APPENDIX A:



### STATA & R ANALYSIS

This appendix provides a detailed statistical analysis and results derived from Stata and 'R'. Analysis is included as an embedded object in the supplementary document and can be accessed here. Double-click the icon to view the document or open it in the browser,

<https://docs.google.com/document/d/1loogVpekeZLnorQyfBGn4ucyQttUpDzE/edit>

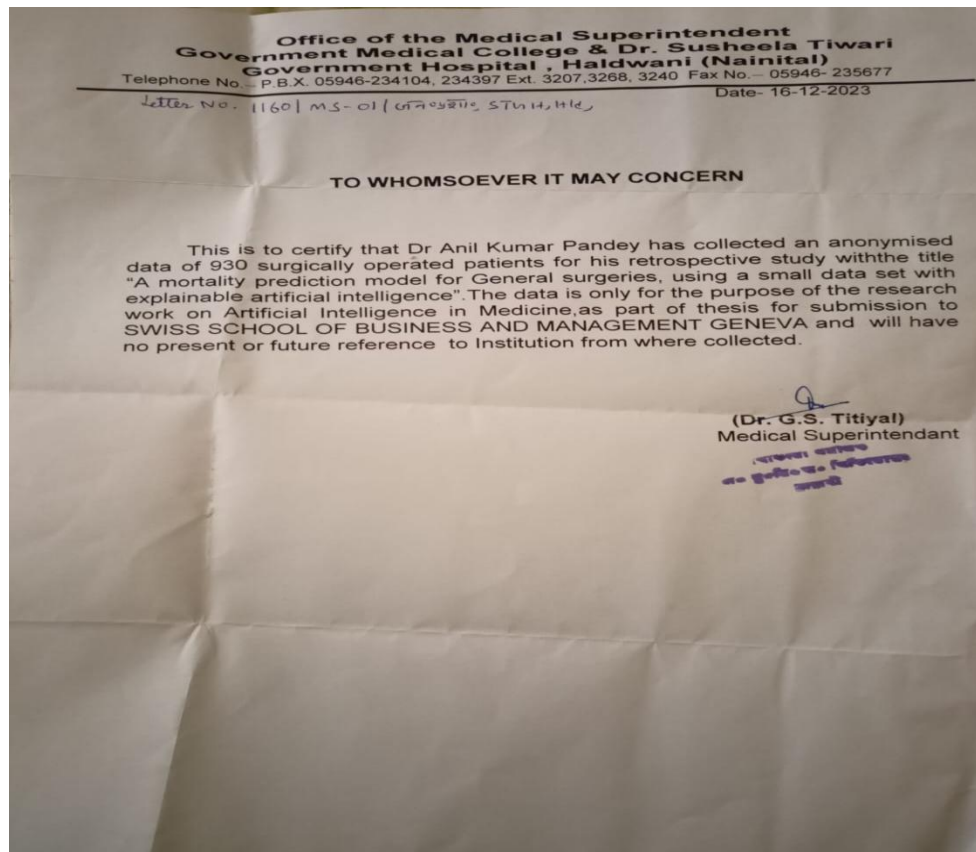
**APPENDIX B**  
**INFORMED CONSENT**



Consent Proof for participation in research

---

Research project title: Mortality Prediction Model for General Surgeries Using a Small data set with Explainable Artificial Intelligence  
Research investigator: Dr. Anil Kumar Pandey  
Phone Number: 9810986268  
E-Mail: [akpandey\\_in@hotmail.com](mailto:akpandey_in@hotmail.com)



## APPENDIX C



Jupyter Notebook for Model Implementation, Calibration, Evaluation and Final model Selection for prediction of Mortality,

In order to develop a well-calibrated model with high predictive efficacy in the first step VAE, Probabilistic models with flipout in last layer and flipout in all layer and a Bayesian model were developed followed by a second step where the ensembles were developed and were calibrated with Temp Scaling, Platt's scaling and Isotonic regression. Model were evaluated to find the best model for prediction of mortality.

### **Link to Jupyter Notebook for Model Implementation and Results**

[!\[\]\(950a62bbddad88d64435fd35607dfc42\_img.jpg\) New\\_research\\_Sep\\_22\\_final\\_28\\_after\\_bty\(2\).ipynb](#)

[!\[\]\(5a132f13505a6571904d622757b7a8f0\_img.jpg\) models\\_1\\_2.html](#)

[!\[\]\(10f8862fc183b400327470ea85afe9ae\_img.jpg\) models\\_1.html](#)


## APPENDIX D



### MODEL INTERPRETABILITY USING LIME AND SHAP ANALYSIS

This appendix provides insights into model interpretability using Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). Both techniques were used with the trained models to understand feature importance and model behavior for individual predictions.

#### **Link to Jupyter Notebook for Model interpretability and Results using LIME**

 [interpretable\\_explainable\\_Final\\_24\\_22.5.pdf](#)

#### **Link to Jupyter Notebook for Model interpretability and Results using SHAP**

 [lime\\_new\\_May\(1\)\(1\).pdf](#)

## REFERENCES

- Ahmed, F.S., Ali, L., Joseph, B.A., Ikram, A., Mustafa, R.U. and Bukhari, S.A., 2020. A statistically rigorous deep neural network approach to predict mortality in trauma patients admitted to the intensive care unit. *Journal of Trauma and Acute Care Surgery*, 89(4), pp.736-742.
- Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M. and Farhan, L., 2021. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1), pp.1-74.
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D., 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
- Austin, S.R., Wong, Y.N., Uzzo, R.G., Beck, J.R. and Egleston, B.L., 2015. Why do summary comorbidity measures such as the Charlson comorbidity index and Elixhauser score work? *Medical Care*, 53(9), pp.e65.
- Bandstra, M.S., Curtis, J.C., Ghawaly Jr, J.M., Jones, A.C. and Joshi, T.H., 2023. Explaining machine-learning models for gamma-ray detection and identification. *Plos One*, 18(6), p.e0286829.
- Bannay, A., Chaignot, C., Blotière, P.O., Basson, M., Weill, A., Ricordeau, P. and Alla, F., 2016. The best use of the Charlson comorbidity index with electronic health care database to predict mortality. *Medical Care*, 54(2), pp.188-94.
- Barash, Y., Soffer, S., Grossman, E., Tau, N., Sorin, V., BenDavid, E., Irony, A., Konen, E., Zimlichman, E. and Klang, E., 2022. Alerting on mortality among patients discharged from the emergency department: a machine learning model. *Postgraduate Medical Journal*, 98(1157), pp.166-71.



Bengio, Y. (2012) 'Practical recommendations for gradient-based training of deep architectures', *Neural Networks: Tricks of the Trade*, Springer, Berlin, pp. 437–478.

Bengio, Y., Courville, A. and Vincent, P., 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), pp.1798-1828.

Berlinet, A. and Thomas-Agnan, C., 2011. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Bernard, S., Heutte, L. and Adam, S., 2009. Influence of hyperparameters on random forest accuracy. In *Multiple Classifier Systems: 8th International Workshop, MCS 2009, Reykjavik, Iceland, June 10-12, 2009. Proceedings* (pp. 171-180). Springer Berlin Heidelberg.

Bilimoria, K.Y., Liu, Y., Paruch, J.L., Zhou, L., Kmiecik, T.E., Ko, C.Y. and Cohen, M.E., 2013. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *Journal of the American College of Surgeons*, 217(5), pp.833-42.

Bishop, C.M. and Nasrabadi, N.M., 2006. *Pattern recognition and machine learning*. New York: Springer.

Blundell, C., Cornebise, J., Kavukcuoglu, K. and Wierstra, D., 2015. Weight uncertainty in neural networks. In *International Conference on Machine Learning* (pp. 1613-1622). PMLR.

Boehmke, B. and Greenwell, B.M., 2019. *Hands-on machine learning with R*. CRC Press.

Bratko, I., 1997. *Machine learning: Between accuracy and interpretability*. Springer Vienna.

- Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., 1984. *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks.
- Breiman, L., 2001. Statistical modelling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), pp.199-231.
- Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp.5-32.
- Briongos-Figuero, L.S., Cobos-Siles, M., Gabella-Martín, M., Abadía-Otero, J., Lobo-Valentin, R., Aguado-De-La-Fuente, A., Vargas-Ruiz, B. and Martín-Escudero, J.C., 2020. Evaluation and characterization of multimorbidity profiles, resource consumption and healthcare needs in extremely elderly people. *International Journal for Quality in Health Care*, 32(4), pp.266-70.
- Brooks, M.J., Sutton, R. and Sarin, S., 2005. Comparison of Surgical Risk Score, POSSUM and p-POSSUM in higher-risk surgical patients. *Journal of British Surgery*, 92(10), pp.1288-92.
- Cabitza, F., Rasoini, R. and Gensini, G.F., 2017. Unintended consequences of machine learning in medicine. *Jama*, 318(6), pp.517-8.
- Caruana, R., 2004. Predicting good probabilities with supervised learning. In *Proceedings of NIPS 2004 Workshop on Calibration and Probabilistic Prediction in Supervised Learning*.
- Castelvecchi, D., 2016. Can we open the black box of AI? *Nature News*, 538(7623), p.20.
- Charlson, M., Szatrowski, T.P., Peterson, J. and Gold, J., 1994. Validation of a combined comorbidity index. *Journal of Clinical Epidemiology*, 47(11), pp.1245-51.
- Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, pp.321-357.

Chen, J.H. and Asch, S.M., 2017. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *The New England Journal of Medicine*, 376(26), pp.2507.

Chen, L., Dubrawski, A., Clermont, G., Hravnak, M. and Pinsky, M.R., 2015. Modelling risk of cardio-respiratory instability as a heterogeneous process. In *AMIA Annual Symposium Proceedings* (Vol. 2015, p. 1841). American Medical Informatics Association.

Chen, P.F., Chen, L., Lin, Y.K., Li, G.H., Lai, F., Lu, C.W., Yang, C.Y., Chen, K.C. and Lin, T.Y., 2022. Predicting postoperative mortality with deep neural networks and natural language processing: model development and validation. *JMIR Medical Informatics*, 10(5), p.e38241.

Chen, T. and Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13 Aug, pp.785-794.

Chiew, C.J., Liu, N., Wong, T.H., Sim, Y.E. and Abdullah, H.R., 2020. Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission. *Annals of Surgery*, 272(6), p.1133.

Chirikov, V.V., Shaya, F.T., Onukwugha, E., Mullins, C.D., dosReis, S. and Howell, C.D., 2017. Tree-based claims algorithm for measuring pretreatment quality of care in Medicare disabled hepatitis C patients. *Medical Care*, 55(12), pp.e104-e112.

Choi, R.Y., Coyner, A.S., Kalpathy-Cramer, J., Chiang, M.F., and Campbell, J.P., 2020. Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science & Technology*, 9(2), p.14. Available: <https://doi.org/10.1167/tvst.9.2.14>

Chollet, F., 2018. *Deep Learning with Python*. Shelter Island, NY: Manning Publications Co.

- Churpek, M.M., Yuen, T.C., Winslow, C., Meltzer, D.O., Kattan, M.W. and Edelson, D.P., 2016. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical Care Medicine*, 44(2), p.368.
- Cohen, M.E., Bilimoria, K.Y., Ko, C.Y., Richards, K. and Hall, B.L., 2009. Effect of subjective preoperative variables on risk-adjusted assessment of hospital morbidity and mortality. *Annals of Surgery*, 249(4), pp.682-689.
- Copeland, G.P., Jones, D. and Walters, M.P., 1991. POSSUM: a scoring system for surgical audit. *British Journal of Surgery*, 78(3), pp.355-360.
- Cosgriff, C.V. and Celi, L.A., 2020. Deep learning for risk assessment: all about automatic feature extraction. *British Journal of Anaesthesia*, 124(2), pp.131-133.
- Dailiana, Z., Papakostidou, I., Varitimidis, S., Michalitsis, S.G., Veloni, A. and Malizos, K.N., 2013. Surgical treatment of hip fractures: factors influencing mortality. *Hippokratia*, 17(3), pp.252.
- Daniel, W. W. (1999). *Biostatistics: A Foundation for Analysis in the Health Sciences* (7th ed.). New York: Wiley.
- Daumé, H., 2009. **Frustratingly easy domain adaptation**. In: *Proceedings of the 47th annual meeting of the ACL and the 4th IJCNLP of the AFNLP*. Suntec, Singapore, 2-7 August 2009. Stroudsburg, PA: Association for Computational Linguistics, pp.256-263.
- Deisenroth, M.P., Faisal, A.A. and Ong, C.S., 2020. *Mathematics for machine learning*. Cambridge University Press.
- Deisenroth, M.P., Fox, D. & Rasmussen, C.E. 2015, 'Gaussian Processes for Data-Efficient Learning in Robotics and Control', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 408–423.

Delahanty, R.J., Kaufman, D. and Jones, S.S., 2018. Development and evaluation of an automated machine learning algorithm for in-hospital mortality risk adjustment among critical care patients. *Critical Care Medicine*, 46(6), pp.e481-e488.

Dempster, A.P., 1968. A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2), pp.205-232.

Dietterich, T.G. (2000) 'Ensemble methods in machine learning', International Workshop on Multiple Classifier Systems, Springer, Berlin, pp. 1–15.

Dominick, K.L., Dudley, T.K., Coffman, C.J. and Bosworth, H.B., 2005. Comparison of three comorbidity measures for predicting health service use in patients with osteoarthritis. *Arthritis Care & Research*, 53(5), pp.666-672.

Drummond, C. and Holte, R.C., 2003. C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II* (Vol. 11, No. 1-8).

Elixhauser, A., Steiner, C., Harris, D.R. and Coffey, R.M., 1998. Comorbidity measures for use with administrative data. *Medical Care*, 36(1), pp.8-27.

Escobar, G.J., Greene, J.D., Scheirer, P., Gardner, M.N., Draper, D. and Kipnis, P., 2008. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. *Medical Care*, 46(3), pp.232-239.

Fajardo, V.A., Findlay, D., Housmanfar, R., Jaiswal, C., Liang, J. and Xie, H., 2018. Vos: a method for variational oversampling of imbalanced data. *arXiv preprint arXiv:1809.02596*.

Fleuren, L.M., Klausch, T.L., Zwager, C.L., Schoonmade, L.J., Guo, T., Roggeveen, L.F., Swart, E.L., Girbes, A.R., Thorat, P., Ercole, A. and Hoogendoorn, M., 2020. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Medicine*, 46(3), pp.383-400.

- Forte, J.C., Wiering, M.A., Bouma, H.R., Geus, F. and Epema, A.H., 2017. Predicting long-term mortality with first-week post-operative data after Coronary Artery Bypass Grafting using Machine Learning models. In *Machine Learning for Healthcare Conference* (pp. 39-58). PMLR.
- Freund, Y. and Schapire, R.E., 1996. Experiments with a new boosting algorithm. In *ICML* (Vol. 96, pp. 148-156).
- Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5), pp.1189-1232.
- Frizzell, J.D., Liang, L., Schulte, P.J., Yancy, C.W., Heidenreich, P.A., Hernandez, A.F., Bhatt, D.L., Fonarow, G.C. and Laskey, W.K., 2017. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA Cardiology*, 2(2), pp.204-209.
- Gagne, J.J., Glynn, R.J., Avorn, J., Levin, R. and Schneeweiss, S., 2011. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *Journal of Clinical Epidemiology*, 64(7), pp.749-759.
- Gal, Y. and Ghahramani, Z., 2016. **Dropout as a Bayesian approximation: Representing model uncertainty in deep learning.** In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML)*, New York, USA, 19-24 June 2016, pp.1050-1059.
- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B., 2004. *Bayesian data analysis*. 2nd ed. Chapman and Hall/CRC.
- Ghahramani, Z. (2016).** Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553), pp. 452–459.

- Glance, L.G., Dick, A.W., Osler, T.M. and Mukamel, D.B., 2006. Accuracy of hospital report cards based on administrative data. *Health Services Research*, 41(4p1), pp.1413-1437.
- Gondara, L., 2016. Medical image denoising using convolutional denoising autoencoders. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* (pp. 241-246). IEEE.
- Gonzalez, J., & Franks, J. (2018). Comparing LIME and SHAP: A Methodological Review of Local Interpretable Machine Learning. In *Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM)*.
- Goodfellow, I., Bengio, Y. and Courville, A., 2016. **Deep learning**. Cambridge, MA: MIT Press.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, pp.2672-2680.
- Goodman, B. and Flaxman, S., 2017. European union regulations on algorithmic decision-making and a right to explanation. *AI Magazine*, 38(3), pp.50–57.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D., 2018. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), pp.93:1–93:42.
- Guillame-Bert, M., Dubrawski, A., Wang, D., Hravnak, M., Clermont, G. and Pinsky, M.R., 2017. Learning temporal rules to forecast instability in continuously monitored patients. *Journal of the American Medical Informatics Association*, 24(1), pp.47-53.
- Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q., 2017. On calibration of modern neural networks. In *International Conference on Machine Learning* (pp. 1321-1330). PMLR.

Hall, P., 2018. *On the Importance of Interpretability in Machine Learning*. arXiv preprint arXiv:1811.02039. Available at: <https://arxiv.org/abs/1811.02039> [Accessed 15 August 2024].

Hall, P., Gill, N., Kurka, M. and Phan, W., 2017. Machine learning interpretability with h2o driverless AI. *H2O.ai*.

Hannan, E.L., Kilburn, H., Lindsey, M.L. and Lewis, R., 1992. Clinical versus administrative databases for CABG surgery: does it matter?. *Medical Care*, 30(10), pp.892-907.

Hastie, T., Tibshirani, R. and Friedman, J.H., 2009. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York: Springer.

Hawkins, D.M. (2004) 'The problem of overfitting', *Journal of Chemical Information and Computer Sciences*, 44(1), pp. 1–12.

He, J., Baxter, S.L., Xu, J., Xu, J., Zhou, X. and Zhang, K. (2020) 'The practical implementation of artificial intelligence technologies in medicine', *Nature Medicine*, 25(1), pp. 30–36. doi:10.1038/s41591-018-0307-0.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. **Deep residual learning for image recognition**. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, USA, 27-30 June 2016, pp.770-778.

Hill, B.L., Brown, R., Gabel, E., Rakocz, N., Lee, C., Cannesson, M., Baldi, P., Loohuis, L.O., Johnson, R., Jew, B. and Maoz, U., 2019. An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. *British Journal of Anaesthesia*, 123(6), pp.877-886.

Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.R. and Samek, W., 2020. *Towards interactive machine learning and the future of artificial intelligence*. *Nature*



*Reviews Computer Science*, 1(9), pp. 441–459. Available at: <https://doi.org/10.1038/s43588-020-00007-1> [Accessed 15 August 2024].

Hu, X., Chu, L., Pei, J., Liu, W. and Bian, J., 2021. Model complexity of deep learning: A survey. *Knowledge and Information Systems*, 63(10), pp.2585-2619.

Huang, Y., Zhang, L., Lian, G., Zhan, R., Xu, R., Huang, Y., Mitra, B., Wu, J. and Luo, G., 2016. A novel mathematical model to predict prognosis of burnt patients based on logistic regression and support vector machine. *Burns*, 42(2), pp.291-299.

Islam, Z., Abdel-Aty, M., Cai, Q. and Yuan, J., 2021. Crash data augmentation using variational autoencoder. *Accident Analysis & Prevention*, 151, p.105950.

Japkowicz, N. and Stephen, S., 2002. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5), pp.429-449.

Joshi, P. and Dhar, R., 2022. EpICC: A Bayesian neural network model with uncertainty correction for a more accurate classification of cancer. *Scientific Reports*, 12(1), p.14628.

Joshi, R. & Dhar, S. (2022). Bayesian neural networks: A practical introduction. *Journal of Artificial Intelligence Research*, 55, 1-34.

Kamthe, S. and Deisenroth, M.P., 2018. Data-efficient reinforcement learning with probabilistic model predictive control. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp.1701-1710.

Kang, D.H., Kim, Y.J., Kim, S.H., Sun, B.J., Kim, D.H., Yun, S.C., Song, J.M., Choo, S.J., Chung, C.H., Song, J.K., Lee, J.W., 2012. Early surgery versus conventional treatment for infective endocarditis. *New England Journal of Medicine*, 366(26), pp.2466-2473.

Kasim, S., Malek, S., Cheen, S., Safiruz, M.S., Ahmad, W.A., Ibrahim, K.S., Aziz, F., Negishi, K. and Ibrahim, N., 2022. In-hospital risk stratification algorithm of Asian elderly patients. *Scientific Reports*, 12(1), p.17592.

- Kendall, A. and Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, pp.5580-5590.
- Kil, S.R., Lee, S.I., Khang, Y.H., Lee, M.S., Kim, H.J., Kim, S.O. and Jo, M.W., 2012. Development and validation of comorbidity index in South Korea. *International Journal for Quality in Health Care*, 24(4), pp.391-402.
- Kingma, D.P. and Welling, M., 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4), pp.307-392.
- Kingma, D.P. and Welling, M., 2014. Stochastic gradient VB and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR 2014* (Vol. 19, p.121).
- Knaus, W.A., Wagner, D.P., Draper, E.A., Zimmerman, J.E., Bergner, M., Bastos, P.G., Sirio, C.A., Murphy, D.J., Lotring, T., Damiano, A. and Harrell, F.E., 1991. The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100(6), pp.1619-1636.
- Krittanawong, C., Virk, H.U., Kumar, A., Aydar, M., Wang, Z., Stewart, M.P. and Halperin, J.L., 2021. Machine learning and deep learning to predict mortality in patients with spontaneous coronary artery dissection. *Scientific Reports*, 11(1), pp.1-10.
- Kuhn, M. and Johnson, K., 2013. Over-fitting and model tuning. *Applied Predictive Modeling*, pp.61-92.
- Lakkaraju, H., Arsov, N. and Bastani, O., 2020. Robust and stable black box explanations. In *International Conference on Machine Learning* (pp. 5628-5638). PMLR.
- Lakshminarayanan, B., Pritzel, A. and Blundell, C., 2017. **Simple and scalable predictive uncertainty estimation using deep ensembles**. In: *Proceedings of the 31st Conference on*

*Neural Information Processing Systems (NIPS)*. Long Beach, USA, 4-9 December 2017, pp.6402-6413.

Le Gall, J.R., Lemeshow, S. and Saulnier, F., 1993. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *Jama*, 270(24), pp.2957-2963.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *Nature*, 521(7553), pp.436-444.

Lee, C.K., Hofer, I., Gabel, E., Baldi, P. and Cannesson, M., 2018. Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. *Anesthesiology*, 129(4), pp.649-662.

Lee, S.W., Lee, H.C., Suh, J., Lee, K.H., Lee, H., Seo, S., Kim, T.K., Lee, S.W. and Kim, Y.J., 2022. Multi-center validation of machine learning model for preoperative prediction of postoperative mortality. *npj Digital Medicine*, 5(1), p.91.

Leeds, I.L., Truta, B., Parian, A.M., Chen, S.Y., Efron, J.E., Gearhart, S.L., Safar, B. and Fang, S.H., 2017. Early surgical intervention for acute ulcerative colitis is associated with improved postoperative outcomes. *Journal of Gastrointestinal Surgery*, 21, pp.1675-1682.

Lieffers, J.R., Baracos, V.E., Winget, M. and Fassbender, K., 2011. A comparison of Charlson and Elixhauser comorbidity measures to predict colorectal cancer survival using administrative health data. *Cancer*, 117(9), pp.1957-1965.

Lien, F., Wang, H.Y., Lu, J.J., Wen, Y.H. and Chiueh, T.S., 2021. Predicting 2-day mortality of thrombocytopenic patients based on clinical laboratory data using machine learning. *Medical Care*, 59(3), pp.245-250.

Lipton, Z.C., 2017. The doctor won't accept that *arXiv preprint arXiv:1711.08037*.

.Lipton, Z.C., 2018. The mythos of model interpretability. *Queue*, 16(3), pp.30:31–30:57.

- Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Liu, F. and Qian, Q., 2022. Cost-Sensitive Variational Autoencoding Classifier for Imbalanced Data Classification. *Algorithms*, 15(5), p.139. <https://doi.org/10.3390/a15050139>
- Lunde, A., Tell, G.S., Pedersen, A.B., Scheike, T.H., Apalset, E.M., Ehrenstein, V. and Sørensen, H.T., 2019. The role of comorbidity in mortality after hip fracture: a nationwide Norwegian study of 38,126 women with hip fracture matched to a general-population comparison cohort. *American Journal of Epidemiology*, 188(2), pp.398-407.
- Martens, D., Huysmans, J., Setiono, R., Vanthienen, J. and Baesens, B., 2008. Rule extraction from support vector machines: an overview of issues and application in credit scoring. *Rule Extraction from Support Vector Machines*, pp.33-63.
- Mi, L., Shen, M. and Zhang, J., 2018. A probe towards understanding GAN and VAE models. *arXiv preprint arXiv:1812.05676*.
- Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities, and challenges. *Briefings in Bioinformatics*, 19(6), 1236-1246.
- Mišić, V.V., Gabel, E., Hofer, I., Rajaram, K. and Mahajan, A., 2020. Machine learning prediction of postoperative emergency department hospital readmission. *Anesthesiology*, 132(5), pp.968-980.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540), pp.529-533.
- Molnar, C. (2020). *Interpretable Machine Learning*. Available online at: <https://christophm.github.io/interpretable-ml-book/>

- Molnar, C., 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. Available at: <https://christophm.github.io/interpretable-ml-book/> [Accessed 15 August 2024].
- Moncada-Torres, A., van Maaren, M.C., Hendriks, M.P., Siesling, S. and Geleijnse, G., 2021. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Scientific Reports*, 11(1), p.6968.
- Moonesinghe, S.R., Mythen, M.G., Das, P., Rowan, K.M. and Grocott, M.P., 2013. Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery: qualitative systematic review. *Anesthesiology*, 119(4), pp.959-981.
- Mueller, S.T., Hoffman, R.R., Clancey, W., and Klein, G., 2019. *Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI*. Tech. Rep., DARPA XAI Program.
- Niculescu-Mizil, A. and Caruana, R., 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pp.625-632.
- Nilsson, J., Ohlsson, M., Thulin, L., Höglund, P., Nashef, S.A. and Brandt, J., 2006. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. *The Journal of Thoracic and Cardiovascular Surgery*, 132(1), pp.12-19.
- Nudel, J., Bishara, A.M., de Geus, S.W., Patil, P., Srinivasan, J., Hess, D.T. and Woodson, J., 2021. Development and validation of machine learning models to predict gastrointestinal leak and venous thromboembolism after weight loss surgery: an analysis of the MBSAQIP database. *Surgical Endoscopy*, 35, pp.182-191.
- Pan, S.J. and Yang, Q. (2010) 'A survey on transfer learning', *IEEE Transactions on Knowledge and Data Engineering*, 22(10), pp. 1345–1359.

Pearse, R.M., Harrison, D.A., James, P., Watson, D., Hinds, C., Rhodes, A., Grounds, R.M. and Bennett, E.D., 2006. Identification and characterisation of the high-risk surgical population in the United Kingdom. *Critical Care*, 10(3), p.R81.

Peng, X., Zhu, T., Wang, T., Wang, F. and Li, K., 2022. Machine learning prediction of postoperative major adverse cardiovascular events in geriatric patients: a prospective cohort study. *BMC Anesthesiology*, 22(1), p.284.

Pera, M., Gibert, J., Gimeno, M., Garsot, E., Eizaguirre, E., Miró, M., Castro, S., Miranda, C., Reka, L., Leturio, S. and González-Duaigües, M., 2022. Machine learning risk prediction model of 90-day mortality after gastrectomy for cancer. *Annals of Surgery*, 276(5), pp.776-783.

Pine, M., Jordan, H.S., Elixhauser, A., Fry, D.E., Hoaglin, D.C., Jones, B., Meimban, R., Warner, D. and Gonzales, J., 2007. Enhancement of claims data to improve risk adjustment of hospital mortality. *Jama*, 297(1), pp.71-76.

Pine, M.I., Jones, B.A. and Lou, Y.B., 1998. Laboratory values improve predictions of hospital mortality. *International Journal for Quality in Health Care*, 10(6), pp.491-501.

**Platt, J.C., 1999.** Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*, MIT Press, pp. 61-74.

Poses, R.M., McClish, D.K., Smith, W.R., Bekes, C. and Scott, W.E., 1996. Prediction of survival of critically ill patients by admission comorbidity. *Journal of Clinical Epidemiology*, 49(7), pp.743-747.

Protopapa, K.L., Simpson, J.C., Smith, N.C. and Moonesinghe, S.R., 2014. Development and validation of the surgical outcome risk tool (SORT). *Journal of British Surgery*, 101(13), pp.1774-1783.

- Quinlan, J.R., 1979. Discovering rules by induction from large collections of examples. In *Expert systems in the microelectronics age*.
- Quinlan, J.R., 1983. Learning efficient classification procedures and their application to chess end games. In *Machine learning: An artificial intelligence approach*, pp.463-482.
- Quinlan, J.R., 1993. *Program for machine learning. C4.5*. Morgan Kaufmann.
- Radford, A., Metz, L. and Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- Rahaman, R., 2021. Uncertainty quantification and deep ensembles. *Advances in Neural Information Processing Systems*, 34, pp.20063-20075.
- Rajkomar, A., Dean, J., & Kohane, I. (2018). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358.
- Rasmussen, C.E. & Williams, C.K.I., 2006. *Gaussian Processes for Machine Learning*. The MIT Press.
- Rau, C.S., Wu, S.C., Chuang, J.F., Huang, C.Y., Liu, H.T., Chien, P.C. and Hsieh, C.H., 2019. Machine learning models of survival prediction in trauma patients. *Journal of Clinical Medicine*, 8(6), p.799.
- Reilly, J.R., Gabbe, B.J., Brown, W.A., Hodgson, C.L. and Myles, P.S., 2021. Systematic review of perioperative mortality risk prediction models for adults undergoing inpatient non-cardiac surgery. *ANZ Journal of Surgery*, 91(5), pp.860-870.
- Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.
- Ribeiro, M. T., Singh, S., & Guestrin, C.** (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144

- Rose, S., 2013. Mortality risk score prediction in an elderly population using machine learning. *American Journal of Epidemiology*, 177(5), pp.443-452.
- Russell, S.J. and Norvig, P., 2016. *Artificial intelligence: a modern approach*. Malaysia: Pearson Education Limited.
- Saklad, M., 1941. Grading of patients for surgical procedures. *Anesthesiology*, 2, pp.281-284.
- Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J. and Müller, K.R., 2021. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3), pp.247-278.
- Schapire, R.E., 2003. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pp.149-171.
- Schölkopf, B., Smola, A.J. and Bach, F., 2002. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press.
- Shapley, L., 1953. A value for n-person games. In *Contributions to the theory of games II*, Princeton University Press, pp.307-317.
- Shawe-Taylor, J. and Cristianini, N., 2004. *Kernel methods for pattern analysis*. Cambridge University Press.
- Singh, R. and Ogunfunmi, T. (2022) 'A comparative study of generative adversarial networks and variational autoencoders for synthetic tabular data generation', *Journal of Big Data*, 9(1), pp. 1-23. doi: 10.1186/s40537-022-00342-0.
- Shilo, S., Rossman, H., & Segal, E. (2020). Axes of AI in medicine. *Cell*, 180(1), 7-10.
- Shorten, C. and Khoshgoftaar, T.M., 2019. **A survey on image data augmentation for deep learning**. *Journal of Big Data*, 6(1), pp.1-48.



- Simard, M., Sirois, C. and Candas, B., 2018. Validation of the combined comorbidity index of Charlson and Elixhauser to predict 30-day mortality across ICD-9 and ICD-10. *Medical Care*, 56(5), pp.441-447.
- Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. and Dieleman, S., 2016. **Mastering the game of Go with deep neural networks and tree search.** *Nature*, 529(7587), pp.484-489.
- Singh, A. and Ogunfunmi, T., 2022. An overview of variational autoencoders for source separation, finance, and bio-signal applications. *Entropy*, 24(1), p.55.
- Smith, D.W., Pine, M., Bailey, R.C., Jones, B., Brewster, A. and Krakauer, H., 1991. Using clinical variables to estimate the risk of patient mortality. *Medical Care*, 29(11), pp.1108-1129.
- Southern, D.A., Quan, H. and Ghali, W.A., 2004. Comparison of the Elixhauser and Charlson/Deyo methods of comorbidity measurement in administrative data. *Medical Care*, pp.355-360.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. **Dropout: A simple way to prevent neural networks from overfitting.** *Journal of Machine Learning Research*, 15(1), pp.1929-1958.
- Steinwart, I. and Christmann, A., 2008. *Support vector machines*. Springer Science & Business Media.
- Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K. and Cilar, L., 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5), p.e1379.

- Stukenborg, G.J., Wagner, D.P. and Connors Jr, A.F., 2001. Comparison of the performance of two comorbidity measures, with and without information from prior hospitalizations. *Medical Care*, pp.727-739.
- Sundararajan, V., Henderson, T., Perry, C., Muggivan, A., Quan, H. and Ghali, W.A., 2004. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *Journal of Clinical Epidemiology*, 57(12), pp.1288-1294.
- Sun, S., Zhang, G., Zhang, C., Yi, C., & Zhang, G. (2019). Variational Bayesian inference for deep learning. *Journal of Machine Learning Research*, 20(8), 1–37.
- Sun, Y., Guo, S., Liu, Y. and Zhang, J., 2019. Bayesian deep learning for uncertainty estimation in medical imaging analysis. *Neurocomputing*, 338, pp.230-241
- Taylor, R.A., Pare, J.R., Venkatesh, A.K., Mowafi, H., Melnick, E.R., Fleischman, W. and Hall, M.K., 2016. Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data–driven, machine learning approach. *Academic Emergency Medicine*, 23(3), pp.269-278.
- Tomescu, V.I., Czibula, G. and Nițică, Ș., 2021. A study on using deep autoencoders for imbalanced binary classification. *Procedia Computer Science*, 192, pp.119-128.
- Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
- Van Walraven, C., Austin, P.C., Jennings, A., Quan, H. and Forster, A.J., 2009. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Medical Care*, pp.626-633.
- Vistisen, S.T., Pollard, T.J., Harris, S. and Lauritsen, S.M., 2022. Artificial intelligence in the clinical setting: Towards actual implementation of reliable outcome predictions. *European Journal of Anaesthesiology*, 39(9), pp.729-732.

Vowels, M.J., 2022. Trying to outrun causality with machine learning: Limitations of model explainability techniques for identifying predictive variables. *arXiv preprint arXiv:2202.09875*.

Vowels, M.J., Camgoz, N.C. and Bowden, R., 2022. *D'ya like DAGs? A survey on structure learning and causal discovery*. *ACM Computing Surveys (CSUR)*, 54(4), pp. 1-36. Available at: <https://doi.org/10.1145/3457603> [Accessed 15 August 2024].

Wallert, J., Tomasoni, M., Madison, G. and Held, C., 2017. Predicting two-year survival versus non-survival after first myocardial infarction using machine learning and Swedish national register data. *BMC Medical Informatics and Decision Making*, 17(1), p.1.

Wang, W., Lee, J., Harrou, F. and Sun, Y., 2020. Early detection of Parkinson's disease using deep learning and machine learning. *IEEE Access*, 8, pp.147635-147646.

Wang, D. & Yeung, D.Y. (2016). Towards Bayesian deep learning: A probabilistic perspective. *Neural Networks*, 88, 36–45.

Wen, Y., Vicol, P., Ba, J., Tran, D. and Grosse, R., 2018. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*.

West, D.M., 2018. *The future of work: robots, AI, and automation*. Brookings Institution Press.

Williams, C.K. and Rasmussen, C.E., 2006. *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.

Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2005. *Data mining: Practical machine learning tools and techniques*. 2nd ed. San Francisco: Morgan Kaufmann.

Wu, J., Roy, J. and Stewart, W.F., 2010. Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches. *Medical Care*, 48(6), pp.S106-S113.

Xu, Y. and Goodacre, R., 2018. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, 2(3), pp.249-262.

Yu, D.T., Black, E., Sands, K.E., Schwartz, J.S., Hibberd, P.L., Graman, P.S., Lanken, P.N., Kahn, K.L., Snyderman, D.R., Parsonnet, J., Moore, R., 2003. Severe sepsis: Variation in resource and therapeutic modality use among academic centers. *Critical Care*, 7(1), p.1.

**Zadrozny, B. & Elkan, C., 2002.** Transforming Classifier Scores into Accurate Probabilities. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02)*, ACM, pp. 694-699.

Zhang, C., Zhou, Y., Chen, Y., Deng, Y., Wang, X., Dong, L. and Wei, H., 2018. Over-sampling algorithm based on VAE in imbalanced classification. In *Cloud Computing – CLOUD 2018: 11th International Conference, Held as Part of the Services Conference Federation, SCF 2018, Seattle, WA, USA, June 25–30, 2018, Proceedings 11*, pp.334-344. Springer International Publishing.

Zhou, Z.H., Wu, J. and Wu, J., 2020. **Machine learning: A practical approach**. Berlin: Springer Nature.