A Thesis

entitled

An Artificial Neural Network Approach to Predict Liver Failure Likelihood

by

Balaji Sathelly

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Master of Science Degree in Engineering: Computer Science

Dr. Ahmad Y. Javaid, Committee Chair

Dr. Scott M. Pappada, Committee Co-Chair

Dr. Mansoor Alam, Committee Member

Dr. Amanda Bryant-Friedrich, Dean
College of Graduate Studies

The University of Toledo
Fall 2018

An Abstract of

An Artificial Neural Network Approach to Predict Liver Failure Likelihood

by

Balaji Sathelly

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Master of Science Degree in Engineering: Computer Science

The University of Toledo
Fall 2018

In recent years, the number of patients with liver disease is rapidly increasing while it remains difficult to detect the symptoms of this disease. A person suffering from liver dysfunction or damage often feels healthy which makes many health care providers fail to detect this condition early on, leading to poor patient outcomes. Such a scenario can be minimized by using clinical decision support systems to optimize detection and prediction of liver failure. Although there are many existing models for liver failure, each of them come with limitations and the issue of liver failure prediction has not been completely resolved to date. In this study, we have addressed this issue by leveraging two comprehensive open-access critical care patient databases to build and validate models for predicting the risk or likelihood of liver failure. Artificial Neural Network (ANN) model architectures that include Multilayer Perceptron (MLP), Generalized Feedforward (GFF), and Modular Neural Network (MNN) were applied to generate a novel 0-100 Liver Failure Risk Index. Models were developed such that an increasing value of the index is associated with an increased risk or likelihood of liver dysfunction. The performance of developed models was compared in terms of sensitivity, specificity, and median lead time for diagnosis. This study has achieved promising results with the best model achieving 83.3% sensitivity at a specificity of 77.5% and correctly diagnosed 83.3%(N = 629 out of 755 possible patients) of liver failure patients. Among these diagnosed patients, the model pre-

dicted the onset of liver failure in 83.5% (N = 525) of patients with a median of 17.5 hours before the onset of liver failure. Hence, our developed models allow health care providers to identify patients at risk of liver failure and facilitate early interventions that may prevent or minimize the associated morbidity and mortality.

This thesis is dedicated to my parents Jagadishwar Sathelly and Padma Sathelly who are the great inspiration and always put my needs ahead of theirs.

Acknowledgments

First and foremost, I would like to express my sheer gratitude and appreciation for my advisor and thesis committee chair Dr. Ahmad Y. Javaid and co-advisor Dr. Scott M. Pappada for all the guidance, leadership, patience, motivation, and support with which they have provided me over the course of my graduate studies. I would also like to thank Dr. Mansoor Alam for agreeing to serve as the member of my thesis committee.

Secondly, I would like to thank The University of Toledo's Department of Electrical Engineering and Computer Science for providing me with financial support in the forms of graduate assistantships and tuition waivers.

Last but not the least, I would like to thank my colleague and friend Mohammad Hamza Owais, for everything he has done. He supported me a lot in tough times during my research. Once again, I would like to thank everyone without whom this would not possible.

Contents

List of Tables

List of Figures

List of Abbreviations

ALD	Alcoholic Liver Disease
ALF	Acute Liver Failure
ALP	Alkaline Phosphate
ALT	Alanine Aminotransferase
ANN	Artificial Neural Network
APACHE II	Acute Physiology and Chronic Health Evaluation II
AST	Aspartate Aminotransferase
AUROC	Area Under Receiver Operating Curve
BMI	Body Mass Index
C5.0	See5
CART	Classification and Regression Trees
CHAID	Chi-Square Automatic Interaction Detector
CTP	Child-Turcotte-Pugh
EMR	Electronic Medical Record
eRI	eICU Research Institute
FPR	False Positive Rate
GFF	Generalized Feedforward
GGT	Gamma-Glutamyltransferase
GSN	Generalized Shunting Neuron
GUI	Graphical User Interface
ICD-9	International Classification of Diseases, Ninth Revision
ICU	Intensive Care Unit
ILPD	Indian Liver Patient Dataset
LD	Lactate Dehydrogenase
LFT	Liver Function Test
LFRI	Liver Failure Risk Index

LM ..... Levenberg Marquardt
LRM ..... Logistics Regression Model
LT ..... Liver Transplantation

MATLAB ..... Matrix Laboratory
MELD ..... Model for End-stage Liver Disease
MELD-Na ..... Model for End-stage Liver Disease-Sodium
MIMIC-III ..... Medical Information Mart for Intensive Care-III
MLP ..... Multi Layer Perceptron
MNN ..... Modular Neural Network

PT ..... Prothrombin Time

ROC ..... Receiver Operating Curve

SD ..... Standard Deviation
SOFA ..... Sequential Organ Failure Assessment

UCI ..... University of California, Irvine

List of Symbols

$\sum$ ..... Summation
® ..... Registered Trademark
ϕ_{LFRI} ..... Threshold value for the LFRI model

Chapter 1

Introduction

Liver failure is a clinical entity characterized by loss of important metabolic and immunological liver functions. Despite significant progress in the last 20 years in the understanding of the pathogenesis of liver failure and the development of management guidelines, this critical illness is still associated with high morbidity and mortality rates of up to 80%- unless a liver transplantation (LT) is performed promptly [1]. An accurate evaluation of the severity of liver failure together with an early identification of its further development is critical in order to determine the further management of the patients. Although liver support devices can be leveraged as temporary treatment, in most cases LT remains the only life-saving treatment of liver failure [2]. LT has been proved to enhance the outcome of these patients by achieving a survival rate of up to 80% [3]. Timely assessment of the likelihood of liver failure is critical for health care providers to make decisions on emergency liver transplantation. As there is a severe shortage of liver donors, it is extremely important to segregate the patients who require LT from the patients who can be treated just by liver support devices. Evaluating whether a patient will require LT or will recover with medical management/treatment is in itself extremely difficult.

1.1 Problem Statement

In most of the cases, LT in liver failure patients could be avoided if the liver failure is detected in early stages. However, detection of a failing liver is complicated in its early stages. Even though health care providers order liver function tests (LFTs) for many people, early detection of this disease remains elusive because abnormal LFTs are indicative of many other diseases besides those related to the liver, such as metastatic malignancy, inflammatory or infective conditioners, and congestive heart failure [4]. Hence, LFTs can be misleading and result in inappropriate treatments leading to increased costs and even morbidity and death [5]. Therefore, an accurate decision support system which can detect the liver failure before its onset is necessary for the proper medication and medical treatment of patients.

1.2 Objective

The functionality of any decision support system depends on the accuracy of its integrated classification and predictive models. The objective of this investigation is to develop a clinically relevant diagnostic and predictive model, which estimates the likelihood or risk of liver failure for a patient in an intensive care unit (ICU) using machine learning, specifically artificial neural networks (ANN). The developed models are designed to output a finite 0-100 liver failure risk index (LFRI). The higher the value of the LFRI, the more likely the patient is to experience liver failure in the future.

1.3 Overview of Liver Functionality

To better understand the severity of liver failure and the importance of identifying its failure well in advance, it is vital to understand liver functionality. The liver is

one of the most important and largest solid organ and gland in the human body that has two sections, called *lobes*. It is situated above, to the right of the stomach (right upper quadrant) and below the diaphragm. The liver plays a key role in keeping the body healthy by detoxifying chemicals and metabolizing drugs. More specifically, its major functions include the production of *bile* which helps the body to absorb fats, proteins, carbohydrates and some vitamins. It also absorbs and metabolizes bilirubin, creates blood-clotting factors (coagulants), metabolizes fats, proteins, and carbohydrates, stores vitamins, and minerals, filters blood, produces albumin, and removes aged and damaged red blood cells [6].

The hepatic artery and portal veins are the two blood sources for the liver; about 2/3 of the blood flow to liver comes from the portal vein and 1/3 from the hepatic artery whereas blood exits via three hepatic veins. The hepatic artery supplies blood and oxygen from heart and lungs to the liver whereas veins supply blood-containing nutrients from the intestine [6].

1.4 Overview of Various Liver Diseases

There are many types of liver diseases that can affect the liver and its functionality. Liver disease, also known as hepatic disease, is a general term and refers to all the potential problems which cause the liver not to perform its designated functions. In general, the functionality of a liver is impacted when at least 75% of its tissue is affected [4]. An early liver disease may have minimal or no symptoms and often will be passed over as being the flu. As liver disease progresses, characteristic signs develop and helps in inferring the cause for it. Based on this cause, liver diseases are classified into many types. Some of these include cirrhosis, viral hepatitis, fatty liver disease, genetic liver disease, and alcoholic liver disease.

Cirrhosis is a condition where liver cells are replaced by fibrous tissue. This

condition can be caused by various factors including alcohol consumption, food contaminated by viruses or bacteria, toxins, and hepatitis. It causes a reduction of the blood flow to the liver which in turn disturbs the functionality of liver [6].

Viral Hepatitis is an inflammation of the liver mainly caused by one of three virus forms- A, B or C. This disease is usually caused by consuming contaminated food or water. Typically, up to 50% of those infected with hepatitis can fight off the virus within six months. However, many patients develop a chronic infection, and the extent of damage to the liver can be determined by a liver biopsy [7].

The fatty liver disease is due to either the excess buildup of fat in the liver or more than 5 to 10% of the total weight of liver. This condition is most commonly seen in people who are diabetic, overweight or have metabolic syndrome. Excess fat in the liver can lead to inflammation and result in cirrhosis in 20% of the patients [8].

Genetic liver diseases are mainly caused due to genetic disorders in patients. The two most common genetic liver diseases are hemochromatosis and alpha-1 antitrypsin deficiency (Alpha -1). Hemochromatosis is the most common adult genetic liver disease in which deposits of iron collect in the liver. Iron deposits may go beyond the liver, affecting other organs such as the heart, joints, and pancreas. Another most commonly seen genetic liver disease in children and adults is Alpha-1 anti-trypsin deficiency. This disease occurs due to the inability to produce enough of a specific protein, called alpha-1 antitrypsin which is used to prevent the breakdown of enzymes in various organs [9].

Alcoholic liver disease (ALD) is one of the major medical complications of alcohol abuse. The three most widely recognized forms of ALD are alcoholic fatty liver (steatosis), acute alcoholic hepatitis, and alcoholic cirrhosis. At least 80% of heavy drinkers develop steatosis, 10%-35% develop alcoholic hepatitis, and approximately 10% develop cirrhosis [10].

1.5 Thesis Organization

This thesis unfolds as follows:

Chapter 2 provides a review of the literature that forms the foundation of this research. It covers a variety of theoretical backgrounds about prior established research methods in the area of early warning scoring systems that have been used and developed to detect the liver failure before its onset.

Chapter 3 gives a brief look at ANN models and the different artificial neural network model architectures used in this study. Then, chapter 4 describes the process of model training and validation set generation, development of neural network models using the generated datasets, and performance analysis and validation of these developed models.

Chapter 4 summarizes the results obtained from the developed models and discusses how this approach overcame the limitations of previous approaches and potential use of the developed model in real-world clinical settings. Chapter 5 draws final remarks and discusses possible future directions in which this research could advance.

Further information and the MATLAB (Matrix Laboratory) source code could be made available upon an email request to either Dr. Ahmad Y. Javaid, Dr. Scott M. Pappada, or Balaji Sathelly.

Chapter 2

Literature Survey

In the past few decades, a number of scoring systems have been used to estimate the likelihood of liver failures such as Sequential Organ Failure Assessment (SOFA), Acute Physiology And Chronic Health Evaluation II (APACHE II), Child-Turcotte-Pugh (CTP), and Model for End-stage Liver Disease (MELD). However, not all of these scores are well validated in liver failure population. Consequently, many new predictive models have recently been developed using techniques such as logistics regression model (LRM), ANN, etc. by considering some common independent predictors like bilirubin total, albumin, prothrombin parameters, etc. These developed models were validated internally (i.e., using a dataset collected at a single institution for model development and validation). Performance of these models was compared with MELD and CTP via evaluating the area under the receiver operating characteristic curves (AUROC). Many of these prior approaches performed better than the previous scoring systems and have shown certain advantages.

2.1 Previous Approaches

For many years, SOFA and APACHE II scores were considered as good indicators of prognosis in critically ill patients. Except for initial scores of more than 11 (mortality rate $> 90\%$), a decreasing SOFA score during the first 48 hours was related

with a mortality rate of less than 6%, while an unchanged or increasing score was related with a mortality rate of 37% when the initial score was 2 to 7 and 60% when the initial score was 8 to 11 in the ICU population [11]. Another scoring system, APACHE II, is a severity-of-disease classification system which is applied within 24 hours of admission of a patient to ICU. Higher APACHE II scores correspond to more severe disease and a higher risk of death. Some of the prior approaches [12, 13] used these scoring systems to predict the likelihood of liver failure in ICU patients. In liver failure patients, APACHE II is superior to SOFA in predicting liver failure and achieved a sensitivity of 66.3% at specificity 76% when it was validated with a total of 725 patients with 300 liver failure and 425 non-liver failure patients [14].

Besides these general scoring systems (SOFA, APACHE II), various early warning scoring systems have been used as the predictive models for liver failure. One of the most widely used models for liver failure is the CTP score [15]. Although CTP was originally used for predicting the mortality during surgery, it is now used to predict the functionality of a liver as well as the necessity of liver transplantation. Another widely used model for assessing the liver failure is MELD. It is calculated from three biochemical variables – creatinine, serum bilirubin, and prothrombin time (PT)- and its performance is more accurate than CTP [16]. To further enhance the performance of MELD, serum sodium (Na) was incorporated into the MELD score, known as MELD-Na, to predict the functionality of liver, especially in cirrhosis patients [17].

Zeng et al. completed an investigation to develop an early warning scoring system to predict the liver failure by using an LRM [18]. This model was developed using a dataset consisting of 242 liver failure and 285 non-liver failure patients and validated with a dataset of 446 (210 liver failure, 236 non-liver failure) ICU patients with the same conditions. This model has identified few independent factors associated with liver failure in ICU patients which include hepatic encephalopathy, hepatorenal syndrome, prothrombin, and age. Performance of this warning system was compared

with MELD and CTP using the AUROC curve and has achieved the highest AUROC of 0.84.

A similar approach was developed by Ren et al. and considered 1000 critical care patients which include 474 liver failure and 526 non-liver failure patients [19]. Training and validation datasets were developed by randomly assigning 60% of both liver failure, and non-liver failure patients to the training dataset and the rest of the patients were used for the validation dataset. These datasets include independent variables like age, white blood cell count, bilirubin total, etc. to develop and validate a multivariate logistic regression model which can be used as an early warning scoring system for liver failure patients in ICU. Moreover, this model's performance was compared with MELD and MELD-Na and achieved a higher AUROC (0.83) than both MELD and MELD-Na.

Many new mathematical models have been developed in recent years to identify liver failure before its onset. For example, Sun et al. developed a novel LRM for identifying liver failure before its onset in 1150 critical care patients (204 liver failure and 946 non-liver failure patients) [20]. The total patients, in this study, were split in the ratio of 70:30 to develop training and validation datasets respectively. The novel LRM developed in this approach included many independent factors such as albumin, prothrombin activity, hepatorenal syndrome, etc. to identify liver failure and compared its performance with the MELD scoring system. The newly established LRM has achieved an AUROC of 0.79 and appears to be superior to the MELD scoring system in estimating the liver failure among the considered 1150 patients.

Another similar approach has been developed to establish a new early warning system for assessing the liver failure risk, named ALPH-Q, which integrates the various clinical and laboratory parameters like age, gender, body mass index (BMI), albumin, total bilirubin, etc. to predict the liver functionality [21]. This approach has considered a total of 874 patients (214 liver failure and 660 non-liver failure) and

generated two datasets by randomly allocating 75% of both types of these patients to training dataset and rest to validation dataset. By using these datasets, ALPH-Q scoring system was developed through Cox Proportional Hazard Regression Analysis, and its performance was compared against CPS, MELD, and previously reported LRM in terms of the AUROC curve. ALPH-Q scoring system achieved an AUROC of 0.83 and performed better than CPS, MELD, and LRM for estimating the liver failure risk in these patients.

Research efforts by Rajanayagam et al. were completed to predict the outcome of the liver failure in children using an ANN approach [22]. In this approach, the ANN model incorporated 34 input variables, compared to 3 input variables required to MELD score. Some of these input variables include alanine aminotransferase (ALT), aspartate aminotransferase (AST), gamma-glutamyltransferase (GGT), albumin, PT, etc. A comprehensive registry-based dataset consisting of 54 children (29 liver failure, 25 non-liver failure) was used to evaluate the predictive outcomes of the developed model and compared its predictive accuracy with MELD using AUROC curves. While AUROC of MELD score was 0.71, ANN model showed a superior performance with AUROC 0.86, sensitivity 82.6%, and specificity 96%.

Further investigation was completed by Moloud et al. to achieve efficient early detection of liver failure through an integrated machine learning modeling approach [23]. This approach included the integration of a multilayer perceptron (MLP) neural network with various decision trees such as see5 (C5.0), chi-square automatic interaction detector (CHAID), and a boosted classification and regression tree (CART). This study has collected 583 records related to the Indian Liver Patient Dataset (ILPD) and 477 non-liver failure patient records from the University of California, Irvine (UCI) repository dataset. This dataset was divided into 70% for the training stage and 30% for the validation stage. Several performance metrics such as sensitivity, specificity, accuracy, etc, were applied in this study. Results indicate that hybridiza-

tion of MLP neural network and C5.0 methods, namely MLPNN-C5.0, achieved the best with a sensitivity of 94.16% and a 93.75% specificity when compared with other algorithms and proved to be a useful approach to diagnose the liver failure before its onset.

Some of the previous approaches were also developed to detect the different types of liver diseases like acute liver failure (ALF) [24], cirrhosis [25], hepatitis A, B and C [26], alcoholic [27], and non-alcoholic related liver diseases [28]. These prior approaches would be undoubtedly useful for health care providers in treating the patients. However, there is a significant need for further research to develop much more effective models and algorithms that support earlier detection or prediction of the onset of liver failure. The identified need motivated our team to develop the predictive models that can diagnose the liver failure in ICU patient population before its onset.

2.2 Advantages of ANN Approach

Physiological systems and its related diseases are extremely complex. Such complex physiological systems and the various parameters and variables that impact and indicate underlying physiological processes can be modeled by using machine learning approaches such as an ANN. ANNs are advantageous for diagnosis and prediction as they consider the effect of variables and parameters which may not be significant by using conventional statistics [29, 30].

Considering the advantages of ANN, many of the previous models [22, 23] were implemented, based on ANN, for diagnosing the liver failure before its onset. Although these approaches have achieved a good predictive capacity, lack of external validation of these developed models and significant size of their validation dataset makes their modeling approaches potentially less significant and limits their clinical utility to a single healthcare institution or a specific group or population of patients.

This implies to a significant need for further research to develop a generalized or universal predictive model which can identify the liver failure before its onset. This model's predictive capacity needs to be significantly high when it is validated with any potential liver failure patient population in representative real-world clinical settings. Developing such a model will serve to assist health care providers in detecting the patients at risk for liver failure and provide a mechanism for earlier treatment and clinical interventions which may provide a means to reduce the morbidity and mortality associated with liver failure and dysfunction.

The summary of the literature survey is presented in Table 2.1. It gives a brief idea of the considered independent factors, implemented technique, performance metrics, and drawbacks of these approaches.

Table 2.1: Recent Early Warning Scoring Systems in Liver Failure

Author	**Features**	**Scoring System or Modeling Technique Used**	**Performance Metrics**	**Drawbacks**
Evangelos et al. [14]	Vital Signs	SOFA, APACHE II	APACHE II score is superior to SOFA, AUROC: Not reported Sensitivity : 66.3% Specificity: 76%	Low sensitivity and specificity

Rahimi et al. [15], Weisner et al. [16], Biggins et al. [17]	Laboratory values	CTP, MELD, MELD-Na	AUROC: CTP– 0.75 MELD- 0.8 MELD-Na – 0.83	Not for ICU patients
Zeng et al. [18]	Hepatic encephalopathy, hepatorenal syndrome, prothrombin, and age	LRM	AUROC: 0.84	Small size and single retrospective studies
Ren et al. [19]	Age, white blood cell count, bilirubin total, etc.	Multivariate logistic regression	AUROC: 0.83 Sensitivity: 88.6% Specificity: 72.2%	Small size and single retrospective studies
Sun et al. [20]	Albumin, prothrombin activity, hepatorenal syndrome, etc.	Novel LRM	AUROC: 0.79	Small size and single retrospective studies

SJ et al. [21]	Age, gender, BMI, albumin, total bilirubin, etc.	Cox proportional hazard regression analysis	AUROC: 0.83 Sensitivity: 79.3% Specificity: 75.0%	Not for ICU population, Small size and single retrospective studies
Rajanayagam et al. [22]	AST, ALT, albumin, GGT, albumin, PT, etc.	ANN	AUROC: 0.86 Sensitivity: 82.6% Specificity: 96%	Not for ICU population, Small size and single retrospective studies
Moloud et al. [23]	Age, gender, albumin, bilirubin, INR, etc.	ANN	AUROC: Not reported Sensitivity: 31.1% Specificity: 95%	Small size and single retrospective studies

Chapter 3

Neural Network Modeling Approach for Predicting Liver Failure Likelihood

Although few of the prior approaches leveraged ANN architectures, many of these well-performed approaches considered statistical techniques like LRM to develop predictive models for estimating the likelihood of liver failure. These statistical techniques allow researchers to develop predictive models which predict the outcome based on the set of independent variables. However, if the researchers include the wrong independent variables, then the model will have little to no predictive value. Also, statistical models are vulnerable to overconfidence and can appear to have more predictive power than they do as a result of sampling bias [31]. All these limitations can be easily overcome by ANN models which can be used to perform nonlinear statistical modeling and provide a new alternative to statistical techniques.

As being a nonlinear statistical data modeling tool, ANNs can reveal the unknown and weak relationships between the input variables and outcome by considering outliers and non-linear interactions among all the existing variables. This ability of neural networks to detect all possible interactions and the availability of multiple training

algorithms motivated us to choose ANN architectures for our study [32].

3.1 A Brief Look at Artificial Neural Networks

ANNs are computer architectures which are modeled after brains. It is built by a series of "neurons" (or "nodes") which are organized in layers [33]. These neurons exhibit global behavior determined by the established connections between the various processing elements and the related parameters within the neural network architecture. Each connection which connects the neurons in consecutive layers is weighted. The weight w_{ij} represents the strength of the connection between i^{th} neuron in a layer and j^{th} neuron in the next layer of the network. The structure of a neural network consists of one "input" layer, one or more "hidden" layers, and one "output" layer. The number of hidden layers and the number of neurons in each of these layers depend on the complexity of the considered system. Figure 3-1 shows a typical ANN architecture with two hidden layers.

In an ANN, data is received through the input layer neurons and then transformed to the neurons in the first hidden layer through the weighted connections established between the input layer and the first hidden layer. Here, the data in each layer are mathematically processed and then the result is transformed to the next layer. Below steps explain how the incoming data (x_i) is processed by j^{th} node in the next layer and Figure 3-2 represents this process.

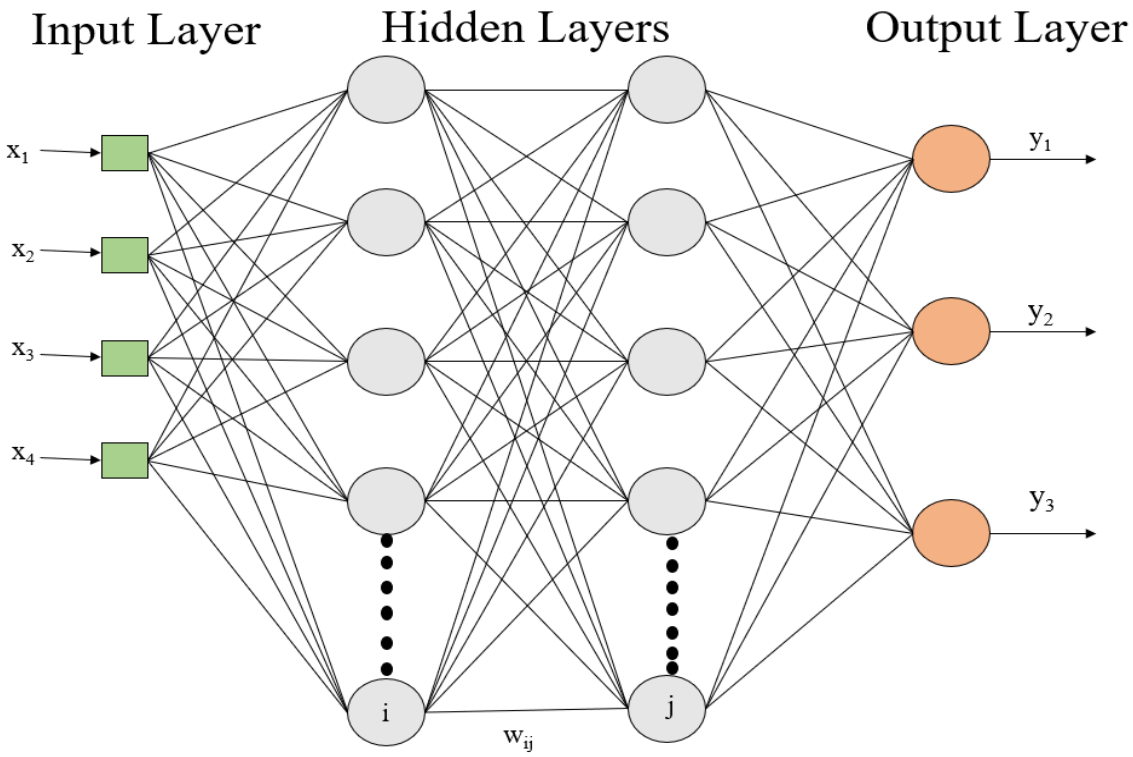

Figure 3-1: General Structure of Artificial Neural Network with Two Hidden Layers

1. First, a weighted sum is calculated and then bias term (θ_j) is added to this sum according to the below equation:

   $net_j = \sum_{i=1}^{m} x_i * w_{ij} + \theta_j (j = 1, 2, ..., n)$ (1)

2. net_j is transformed using a mathematical 'transfer function'. This function is used to normalize all network inputs and outputs to a finite range of values. This allows the neural network to better identify patterns and trends in data better than have a range of values between 50 and 500, for example. Many transfer functions can be used for this. However, a sigmoid function is used in this effort for the reasons summarized in section 'Model Development', and it is shown in the below equation:

   $f(x) = 1 \ / \ (\ 1 + e^{-x} \ )$ (2)

3. Finally, the result is transferred to the neuron in the next layer.

After the development of a neural network to an application, training must be done with the random initial weights chosen. These models are popular because of its adaptive nature and learn by determining patterns existent in input data. Training of a network can be done in two ways - *supervised training* and *unsupervised training*. In supervised training, the desired output should be provided along with the inputs to optimize the network weights to find the best set of weights that lead to minimum error in the output. Some of the areas where models implementing supervised training can be applied are - function approximation, regression analysis, time series prediction, and classification such as pattern and sequence recognition, etc. On the other hand, unsupervised training is applicable when a model must make sense of the inputs on its own and describes the structure of "unlabeled" data, i.e. data which has not been classified. Unsupervised training is used for applications including but not limited to clustering, and anomaly detection [30].

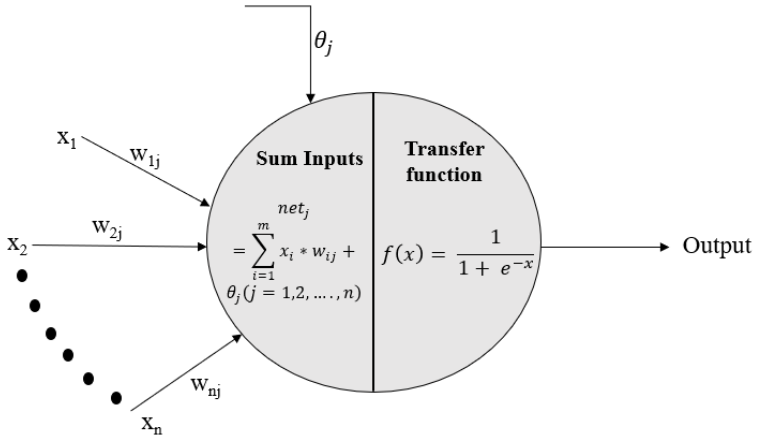

Figure 3-2: Data Processing in a Neuron

3.2 ANN Model Architectures

3.2.1 MultiLayer Perceptron Model

An MLP is a type of feedforward artificial neural network consisting of at least three layers of nodes. Each neuron in both hidden layers and output layer uses a non-linear activation function. Both these multiple layers and non-linear activation function distinguish MLP from standard linear perceptron and helps these networks to differentiate the data which is not linearly separable [34]. MLPs are universal function approximators and can be applied to develop mathematical models using regression analysis. These networks are well suited for a wide variety of modeling applications such as pattern classification, prediction, and function approximation. Pattern classification is concerned with the classification of data into discrete classes. Prediction is related to forecasting of a time series data when the current and previous trends are known whereas function approximation involves the task of modeling the relationship between the variables [35].

3.2.2 Generalized Feedforward ANN Model

A GFF neural network is an ANN where the unit connections do not form a cycle such as recurrent neural network models [36]. This network was the first and the simplest type of ANNs where data moves in only one direction, forward, from the input nodes to the output node through the hidden nodes. This network's architecture uses a *generalized* shunting neuron (GSN) model as its basic computing unit, and this differentiates this network from MLP which is based on perceptrons. These shunting neurons are capable of forming the complex, nonlinear decision boundaries and help the GFF neural network architecture to perform various tasks such as complex pattern classification problems, dynamical modeling, time series forecasting, pattern recognition, and data mining [37, 38].

3.2.3 Modular Neural Network Model

MNN is a particular class of MLP in which several parallel MLPs are used to process the inputs and then recombine the results. This process leads to forming some structure within the topology which helps in developing a specialized function in each sub-module. This approach of *Divide and Conquer* incorporates many advantages to a neural network such as scalability, robustness, flexibility in design, and implementation. Moreover, these networks require a lesser number of weights than an MLP to build a network of similar size because of partial interconnection between its layers. Hence, this reduces the number of required training exemplars and helps in speeding up the training times. However, this network can be segmented into modules in many ways and it is unclear how to best design the modular topology based on the data [39]. Figure 3-3 shows an architecture of a modular neural network with 'k' modules.

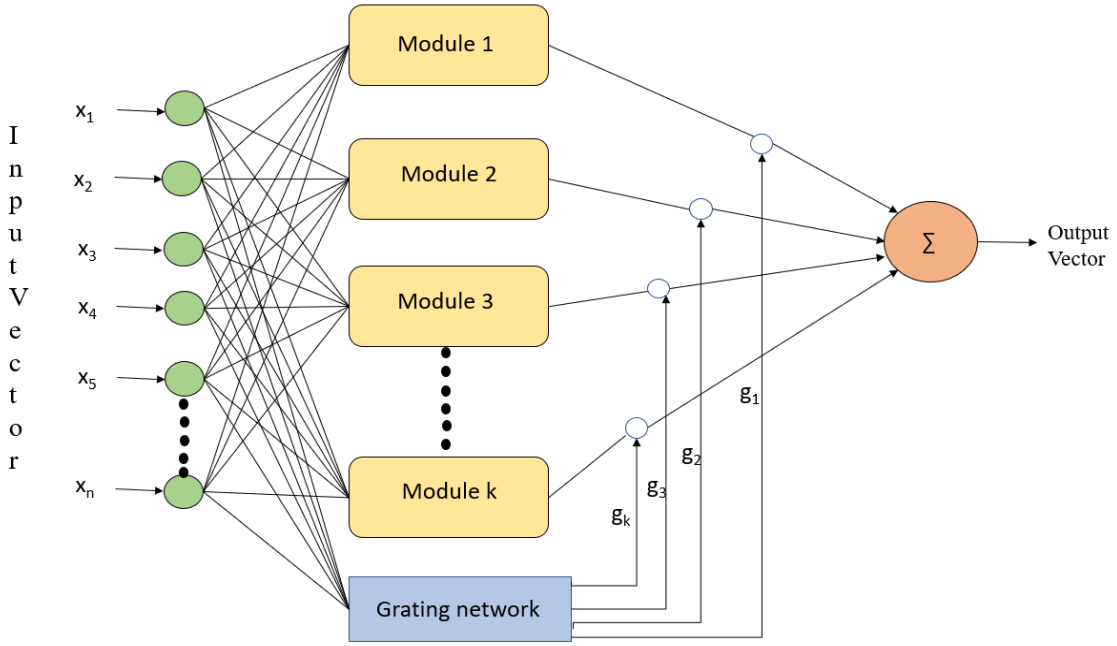

Figure 3-3: Modular Neural Network Architecture

3.3 Model Development and Validation

In this study, we have leveraged the above mentioned ANN architectures for developing multiple models to estimate the likelihood of liver failure in ICU patients. Model development and validation is the process of training a model accurately with a given dataset and validating the performance of this developed model with another dataset. Hence, this process involves two datasets – a training dataset and a validation dataset.

3.3.1 Model Training and Validation Set Generation

To develop and validate the machine learning-based models that are required to generate the LFRI, we leveraged two large open-access critical care databases. The first database used was MIMIC-III (Medical Information Mart for Intensive Care-III). This database is notable as it is freely available to researchers worldwide with a diverse

and very large population of ICU patients. It consists of de-identified health-related information associated with $>$ 45,000 critical care patients admitted from the ICUs of the Beth Israel Deaconess Medical Center from 2001-2012 [40]. This data was used to train the various models aimed at predicting the risk for liver failure discussed in the section below. Model training sets were generated for each ANN using a custom software application we developed in $MATLAB^{®}$ (Mathworks, Natick, MA). The model inputs (categories of input data sources) included in the model training sets for the ANN models consisted of vital signs and laboratory results collected throughout a patient's ICU length of stay. To develop the targeted continuous LFRI (i.e., target model output) for the training set, patient data was evaluated every hour with respect to a set of liver failure diagnostic criterion defined by a collaborating critical care physician. If patients had an ICD-9 (International Classification of Diseases, Ninth revision) diagnosis of liver failure (570 - 573 and its child codes) [41] during their ICU stay and met the clinical diagnostic criteria for liver failure at a given timestamp (evaluated hourly) a "1" was used as the target model output at each time stamp, and this condition held true throughout the patient's ICU stay. Where this condition was not true, a "0" was used as the target or desired model output at each corresponding time stamp.

To assess and validate model performance and accuracy, we used a second open-access database, the eICU Collaborative Research Database [42]. This database is developed through the work of Philips Healthcare and collaborators at the MIT laboratory for computational physiology and maintained by Philips eICU Research Institute (eRI). This database includes time-stamped ICD-9 diagnoses of liver failure and provided the ability to evaluate the accuracy of developed models in both the detection (i.e., diagnosis) and prediction (i.e., predictive diagnosis) of the onset of liver failure. The developed models were validated using 81,135 patients where only 755 patients had a diagnosis of liver failure. Patients were eliminated from the vali-

dation/testing set if they had a diagnosis of liver failure < four hours into their ICU admission (i.e., considered a preexisting diagnosis).

It is important to note and recognize that one of the primary issues facing by researchers in developing machine learning-based models is *missing data*. This is a very common problem seen in most of the retrospective studies using health records databases. Missing data occurs because of the infrequent availability of certain data sources in the dataset. For example, vital signs are taken frequently in the ICU, approximately for every hour whereas certain laboratory results are taken much less frequently, sometimes only once in a day. For our model development, missing data in both model training and validation sets (described above) were handled by replacing it with the last known laboratory value. For example, consider a laboratory data source whose value is missing at a time stamp 't'. We try to replace this missing value with a value of the same data source at timestamp 't-1'. If we could not find a value at this time stamp, then we look for a value at timestamp 't-2' and so on. Finally, if we do not find any laboratory value in the previous instances for this data source, then we consider this missing value as '-1'. Hence, this approach provided the capability to generate a significant dataset for models to develop and validate the models developed during this study using data collected every hour during each patient's ICU length of stay.

3.3.2 Model Development

During this study, we investigated the development and application of three different types of neural network model architectures – MLP, GNN, and MNN. For each of the model architectures investigated, we iteratively changed different model parameters. Some of the common parameters that were considered to change iteratively include – the number of hidden layers, and the number of processing elements in each of these hidden layers. Model development was initiated by building simple

neural networks models and then iteratively increased the complexity of these models by changing the various aforementioned parameters. For example, initially we have built models with two hidden layers and then tried with three hidden layers and so on. Also, we have tried to increase the number of processing elements for each of these hidden layers starting from 5 to 50 to develop different models. Finally, all of these developed models were validated with the generated validation dataset and compared the performances of these models to obtain the best model. In this study, the best models were obtained with two hidden layers having 10, 2 processing elements in the first and second hidden layers respectively.

In summary, this iterative process of changing different parameters to obtain the best performing model involves a lot of computation cost, and manual effort. This process was simplified by using *NeuroSolutions*® software (Neurodimension, Gainesville, FL) which provides an intuitive graphical user interface (GUI) application to support model development and validation efforts. Using this software it is a seamless process to select a desired network architecture and modify various parameters within a model architecture.

All models developed during this study were trained using the *Levenberg Marquardt (LM)* algorithm. This is a backpropagation training algorithm which is more powerful than the conventional gradient descent algorithm. This algorithm is an iterative technique that finds the minimum of a multivariate function which is expressed in terms of the sum of squares of nonlinear least-squares problems. In simple terms, the functionality of this algorithm can be assumed as a combination of steepest descent and the Gaussian-Newton method. When the error is high, LM behaves like a steepest descent method which is a slow technique but converges. When the error is low, it works like gaussian-newton method. This algorithm follows an iterative approach to minimize the error and based on this error's magnitude; training determines the degree of weights adjustment to reduce the overall error of the model [43].

Moreover, these models were configured with *batch* training and designed to terminate the training process when the model has completed 1000 epochs, or the mean square error for the cross-validation dataset does not improve after a set number of epochs or starts increasing which indicates the overfitting of the model.

As described in 'Model Training and Validation Set Generation' section, all ANN models were designed to derive a 0-100 LFRI based on hourly laboratory and vital signs data. All models implement a *sigmoid* transfer function to constrain/normalize model inputs and outputs to a 0-1 range. This function is critical to the design of our approach as this effectively generates an LFRI value ranging between 0 and 1. Based on this design, values closer to 0 indicate that the patient is not trending towards a liver failure state, and values closer to 1 indicate a higher probability that the patient is in a state of liver failure or liver dysfunction or trending towards one. Finally, the model output represents a class membership (i.e., with or without liver failure) and multiplying this output by 100 results in the predetermined 0-100 LFRI value. Based on this model's design, a value of LFRI > 50 would indicate that there is a higher probability that the patient would be at risk for liver failure or dysfunction.

3.3.3 Model Performance Analysis and Validation

Sensitivity and specificity of the LFRI in diagnosing and predicting the onset of liver failure was evaluated using a static model output *threshold value (ϕ_{LFRI})* of 50 based on the model's design justification provided previously. In medical practice, it is nearly impossible to achieve perfect discrimination between diseased patients and healthy patients with a single threshold value (ϕ_{LFRI}). This scenario implies to select the best compromise between sensitivity and specificity by considering different diagnostic test results. As such, AUROC curves should be leveraged to provide a clearer understanding of a model's diagnostic capabilities [44]. The AUROC is the most commonly used receiver operating characteristic (ROC) metric which summa-

rizes the overall diagnostic accuracy of a test or model/classifier. Its value ranges from '0' to '1' where '1' implies that the test is entirely accurate and '0' implies that the test is entirely inaccurate. Hence, the higher the value of AUROC, the better the model's diagnostic capabilities [45]. For this study, the AUROC was calculated to evaluate the diagnostic capabilities of the LFRI with variable thresholds ranging from 0 to 100. Additional model performance metrics included: the *predictive capacity* of the LFRI or % of patients that were correctly diagnosed with liver failure by the LFRI model before the onset of liver failure. We also calculated the mean lead time to diagnosis +/- the standard deviation (SD) and median lead time to diagnosis for each model. The mean lead time is calculated as the average lead time to diagnosis of all patients where liver failure was predicted by the model before a clinical diagnosis was made and SD measures the dispersion of lead times of all these patients relative to its mean. The median lead time represents the central lead time in the group of these sorted lead times obtained for all patients who were detected before clinical diagnosis. For the distribution of lead times, if both the measures - mean and median are significantly different, then it indicates that the distribution of lead times data is skewed, i.e. the data is far from being normally distributed. For such kind of data distribution, the median gives a more appropriate idea about data distribution [46].

Figure 3-4 explains the whole process of neural network model development using different architectures and validation for analyzing its performance.

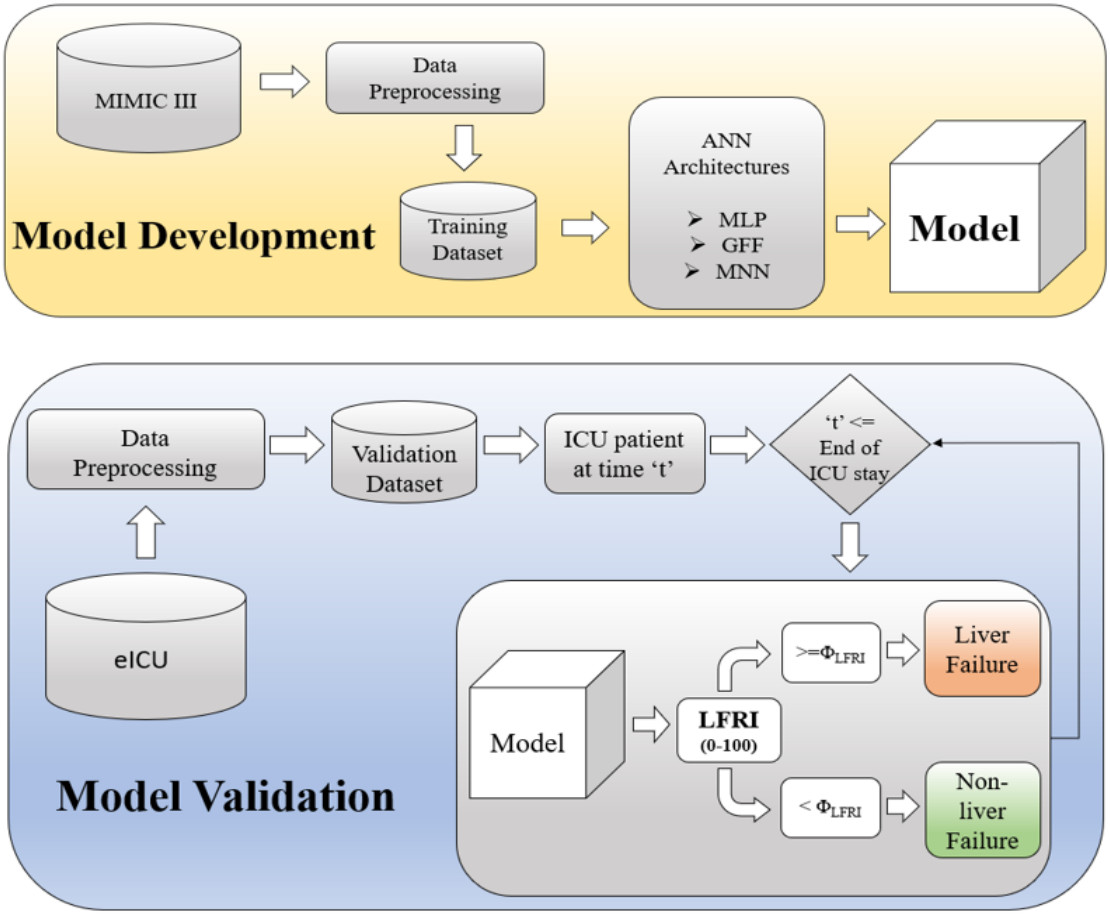

Figure 3-4: Process of Model Development and Validation

Chapter 4

Results and Discussion

In this study, we have validated the developed models against different performance metrics such as sensitivity, specificity, predictive capacity, AUROC, mean, and median lead time. Below description shows a comparative analysis of various developed models in terms of these performance metrics and LFRI plots obtained for the best model.

4.1 Results

As explained in the 'Model Development' section, we have developed various models by iteratively changing different parameters of neural network architectures and validated these developed models with eICU validation dataset. Table 4.1 shows the performance metrics - *sensitivity* and *specificity* - obtained for some of the developed models which were configured with two, three, and four hidden layers and six, four, and three processing elements in each of these hidden layer respectively.

Table 4.1: Performance Metrics

Model	2 Hidden Layers 6 PEs in each hidden layer		3 Hidden Layers 4 PEs in each hidden layer		4 Hidden Layers 3 PEs in each hidden layer	
	Sensitivity	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**
MLP	75.2%	71.5%	72.7%	65.3%	68.1%	64.2%
GFF	69.5%	70.8%	65.2%	66.7%	63.6%	66.3%
MNN	76.8%	35.6%	74.8%	33.2%	72.4%	29.7%

Table 4.2: Confusion Matrix for the best models

Model	**True Positive**	**False Positive**	**True Negative**	**False Negative**
MLP	629	18084	62296	126
GFF	682	45877	34503	61
MNN	578	14734	65646	181

Table 4.2 summarizes the confusion matrix and Table 4.3 includes the model performance metrics calculated for the best models developed by configuring the various neural network model architectures with two hidden layers and 10, 2 processing elements in the first and second hidden layers respectively.

Table 4.3: Performance Metrics for the best models

Model	**Sensitivity**	**Specificity**	**Predictive Capacity**
MLP	83.3%	77.5%	83.5% (N=525)
GFF	76.1%	81.6%	80.1% (N=546)
MNN	91.8%	42.9%	90.3% (N=522)

Figures 4-1, 4-2, and 4-3 contain the ROC curves obtained for the best models and Table 4.4 shows the calculated AUROC for these models. In this investigation, a patient was identified with probable liver failure if his or her LFRI exceeds the specified threshold value (ϕ_{LFRI}). Among the 81,135 patients (755 positives, 80,380 negatives) in the validation set where the final outcome was known, the MLP model, achieved the highest AUROC (as shown in Table 4.4), identified the patients before the onset of liver failure with an AUROC of 0.86 and achieved a sensitivity of 83.3% at a specificity of 77.5%. For all the patients who were detected before clinical diagnosis, we have obtained a distribution of lead times with a mean of 34.4 hours +/- 27.7 hours and median of 17.5 hours. As this distribution was positively skewed (i.e. median << mean), the median would be a better measure of predictive performance of the developed model.

Table 4.4: AUROC for the best models

Model	**AUC**
MLP	0.8622
GFF	0.8618
MNN	0.8266

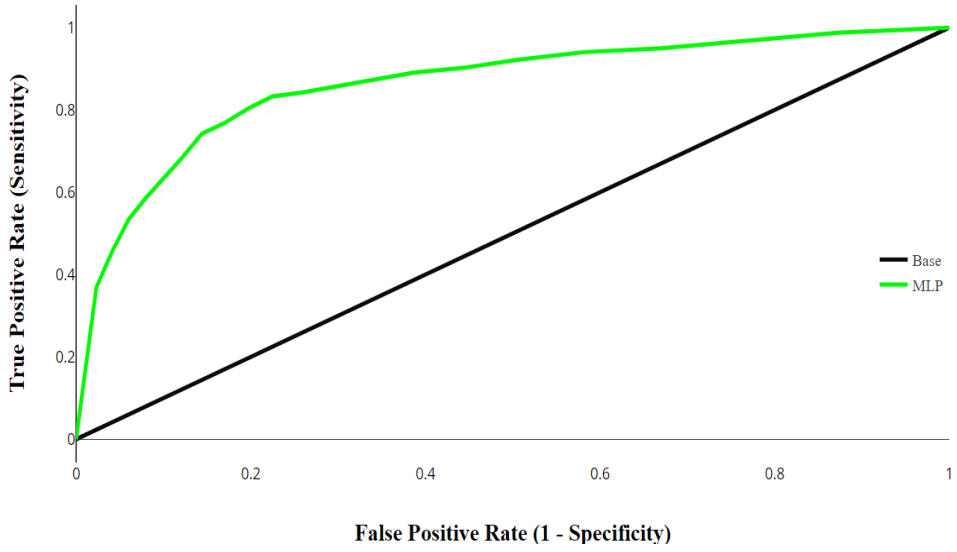

Figure 4-1: ROC Curve for MLP Model

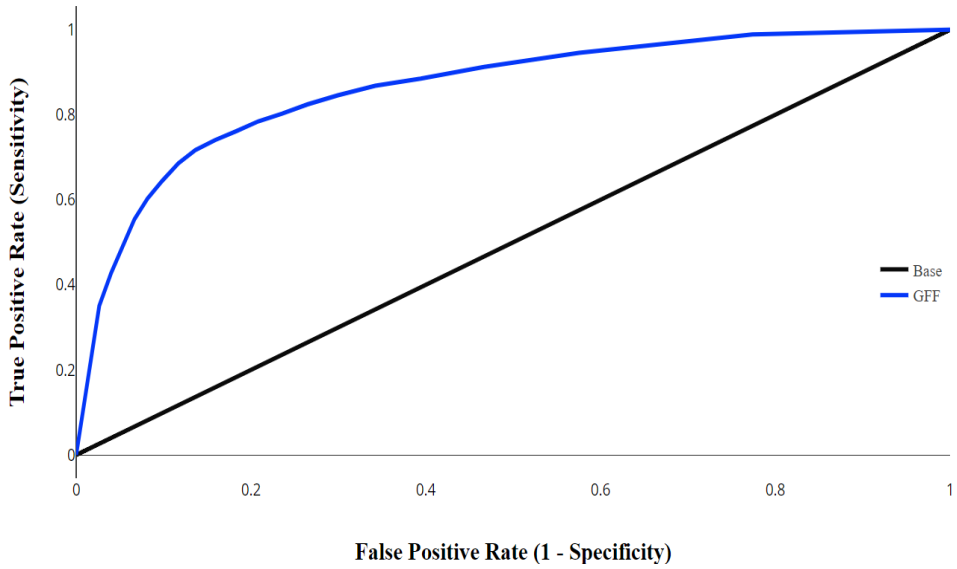

Figure 4-2: ROC Curve for GFF Model

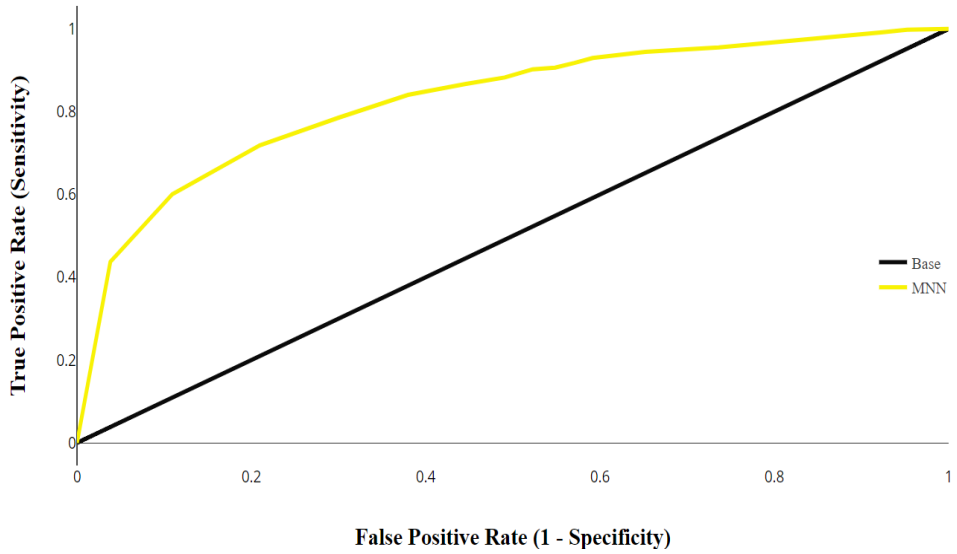

Figure 4-3: ROC Curve for MNN Model

The developed LFRI is intended to alert for impending or concurrent liver failure in a patient when the LFRI reaches or exceeds a threshold value (ϕ_{LFRI}) of 50. Figures 4-4, 4-5, 4-6, 4-7, and 4-8 show the model-generated 0-100 LFRI plots for sample ICU patients from the model validation set. In these figures, the horizontal dotted black line represents the static LFRI threshold value (ϕ_{LFRI}) of 50 implemented for this effort. The circular indicator, in figures 4-4, 4-6, and 4-8 indicate the initial time at which model-generated LFRI exceeded the defined threshold value (ϕ_{LFRI}). The vertical dashed red line in figures 4-4, 4-6, and 4-7 represent the timestamp of the occurrence of an initial clinical diagnosis of liver failure during the patient's ICU length of stay.

Figure 4-4 shows the model-generated 0-100 LFRI plot for a sample liver failure patient who was diagnosed by the model before the onset of liver failure (True Positive). Here, the LFRI exceeded the chosen threshold value (ϕ_{LFRI}) around 70 hours before the onset of liver failure. Also, the LFRI stays above the threshold and continues to increase for the remainder of the patient's ICU admission which indicates that the model has detected many hours before that the patient is more likely to

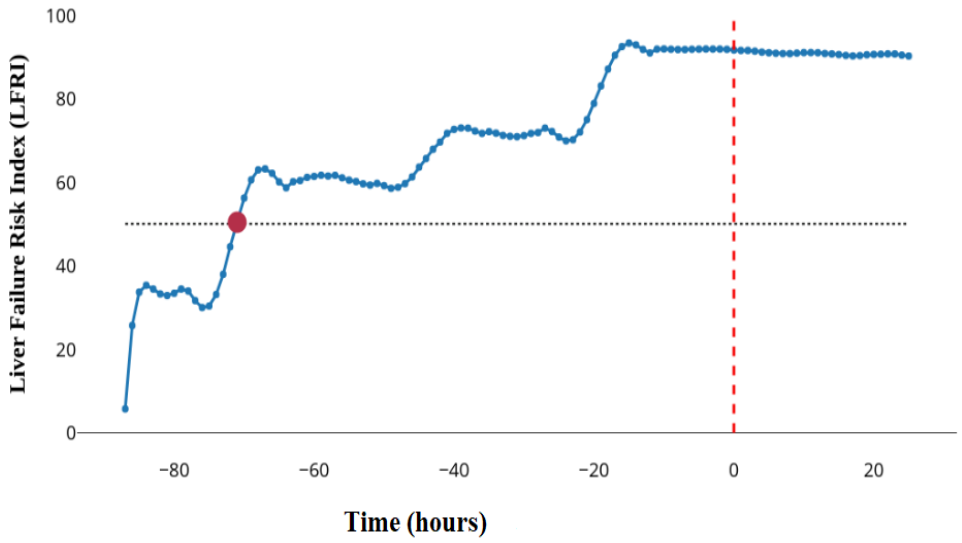

Figure 4-4: Risk Index Plot for Liver Failure Patient A (True Positive)

experience liver failure.

Figure 4-5 shows the model-generated 0-100 LFRI plot for a sample non-liver failure patient (True Negative) from the model validation set. Here, the LFRI is below the threshold value (ϕ_{LFRI}), as desired, during the entire stay of ICU patient which indicates that the patient is less likely to have or experience liver failure.

Figure 4-6 shows a similar LFRI plot, as in Figure 4-4, for a different liver failure patient (True Positive Diagnosis but non-predictive). However, the model diagnosed the patient around 10 hours after the onset of liver failure. This delay could be occurred due to multiple reasons. One potential reason for this delay could be associated with missing data where some of the critical model predictors indicating liver failure or increased risk for liver failure are unavailable in the patient record. As explained in the 'Model Training and Validation Set Generation' section, this data was generated using the last known laboratory values, and if it was in the normal range, then the model cannot detect in the early stages and leads to delay in diagnosis.

Figure 4-7 shows a model-generated 0-100 LFRI plot for a sample liver failure patient where the model-generated LFRI never exceeded the threshold value (ϕ_{LFRI})

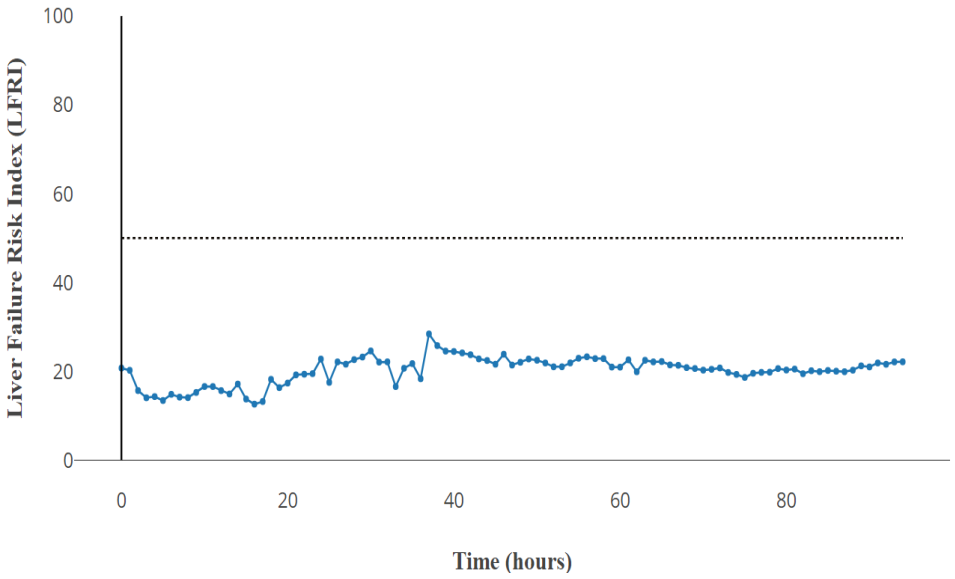

Figure 4-5: Risk Index Plot for Non-Liver Failure Patient B (True Negative)

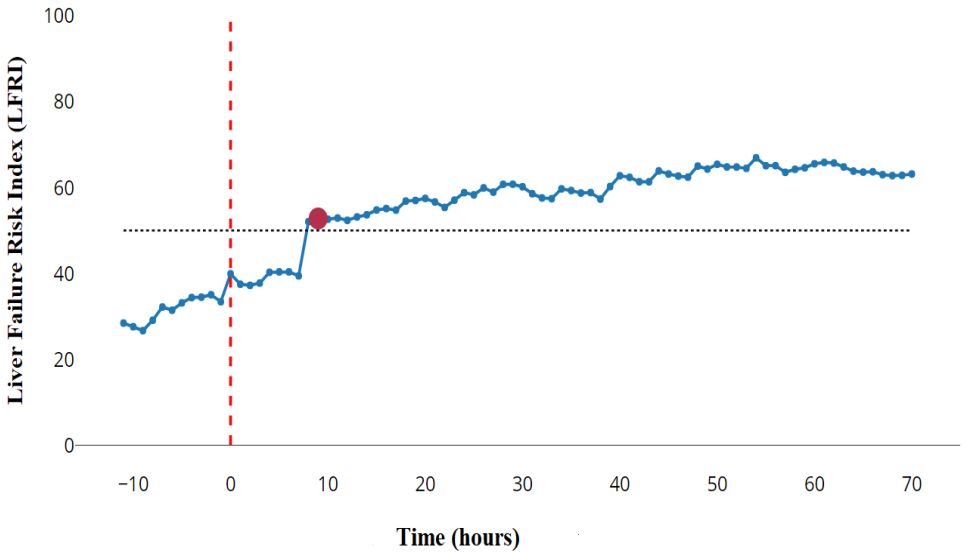

Figure 4-6: Risk Index Plot for Liver Failure Patient C (True Positive with Negative Lead Time)

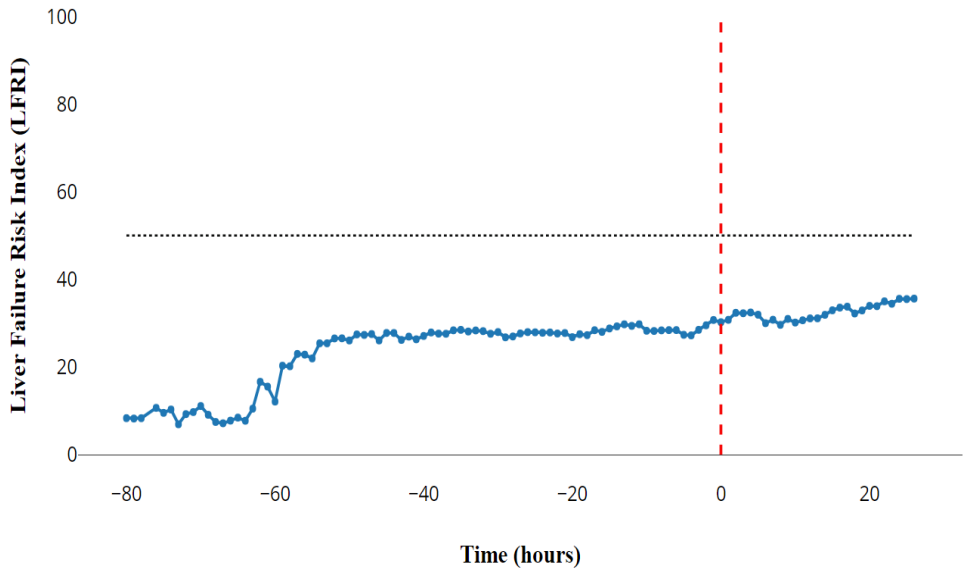

Figure 4-7: Risk Index Plot for Liver Failure Patient D (False Negative)

for this patient and hence wrongly classified as non-liver failure patient (False Negative). This could be due to the absence of some potential LFTs results such as alkaline phosphate (ALP) test as well as the unavailability of other laboratory test results that indicate the liver function status [47]. These LFTs were not incorporated in the databases considered in this study. Including such potential LFTs along with current predictors will help the model to generate a better LFRI and minimizes the false negative cases.

Figure 4-8 shows a model-generated 0-100 LFRI plot for a sample non-liver failure patient. For this patient, although the LFRI was slightly above the threshold value (ϕ_{LFRI}) only once during the entire ICU stay, the model has classified the patient as likely to have a liver failure (i.e., a false positive). This situation could be improved in the future by modifying the alerting algorithm and its corresponding logic for the LFRI model. Potential improvements could be made such as altering the detection criteria such that a patient would be diagnosed only after the LFRI remains above the detection threshold value (ϕ_{LFRI}) for at least 2 hours or some other desired length of time.

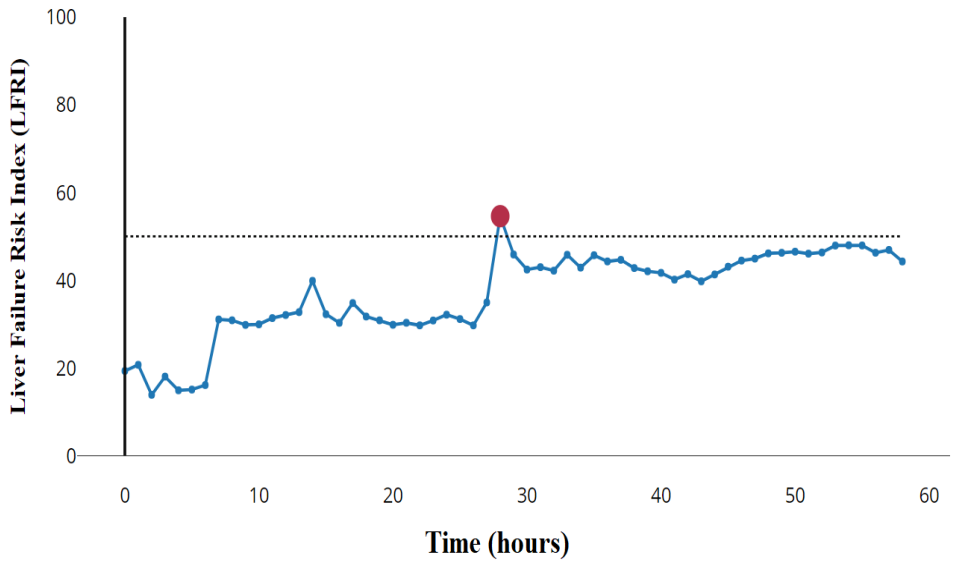

Figure 4-8: Risk Index Plot for Non-Liver Failure Patient E (False Positive)

4.2 Discussion

4.2.1 Summary

Detection of liver failure and dysfunction in its early stages may promote more timely treatment and interventions by health care providers and consequently, enhance patient outcomes. To achieve this, our study has developed a neural network modeling approach to predict the likelihood of liver failure before its onset in ICU patients. As the neural network modeling techniques are capable of estimating the impact of various inputs/independent variable on output/dependent variable, the application of this kind of modeling technique towards the prediction of the likelihood of liver failure is more suitable than the other modeling techniques used in prior approaches.

Some of the previous approaches [12, 13] tried to use organ failure scoring systems such as SOFA, APACHE II to evaluate the illness severity amongst liver failure patients. As these scoring systems are intended for broader organ systems, the high

values of these scores indicate that the patient is at critical (without inspection of the various dimensions of the scoring system that contributes to the final composite value). Therefore, these prior approaches potentially fail to identify the patients who are at the highest risk for a *specific acute condition* with high sensitivity and specificity. Low sensitivity diagnostic tools often fail to identify the correct outcome for many patients, and this leads to uncertainty in benefiting from early treatment. We have overcome this drawback by generating a predictive model exclusively for identifying liver failure patients and achieved promising results of 83.3% sensitivity at a specificity of 77.5% and correctly identified 83.3%(N=629) of patients with liver failure that were present in the eICU validation set. The LFRI successfully predicted the onset of liver failure in 83.5% (N=525) of the 629 patients with a median of 17.5 hours before its onset.

In recent years, many prior approaches implemented various machine learning algorithms to develop the predictive models for early identification of the liver failure. Most of these models however, have not been developed and tailored for use in ICU patient population. The ICU is one of the most crucial functioning operational units in a hospital. Each ICU has a different environment that represents the surgical procedure followed by medical specialists. ICU teams comprise of highly skilled intensive care doctors, specialists, and nurses who are skilled in providing care to critically ill patients using specialized, technical and monitoring equipment. Unlike general patients, daily monitoring of ICU patients is necessary because the optimization of patient statuses including but not limited to: hemodynamic, ventilation and nutrition is critical to improving the survival of patients. So, the surveillance and monitoring of ICU patients are extremely important, and the inability to detect or predict liver failure in these patients may lead to catastrophic consequences [48, 49]. As explained in the 'Model Training and Validation Set Generation' section, we have addressed this challenge by developing models from a broad ICU patient population

with and without liver failure.

In each of the prior approaches [18, 19, 20, 21, 22, 23], models were developed and validated with a relatively small sample size from a single-center retrospective studies. Systematic reviews which evaluate this kind of approaches all conclude that such studies have the characteristics of deficiencies in study design, inadequate statistical methodology, and poor reporting [50, 51]. Moreover, this kind of validation makes the developed models only work effectively for a particular health care institution or a small subset or population of patients which impacts its overall clinical utility. It is necessary to see how well a model performs with patients from a different but "plausibly related" population. Therefore, impact studies should not be considered until the robustness and generalizability of the developed model is verified with one or more external validation databases [52].

Further, all the previous approaches [18, 19, 20, 21, 22, 23] developed and validated the predictive models using datasets which included an almost equal number of both liver failure and non-liver failure patients. Testing a developed predictive model with a validation dataset which has an equal ratio of with and without liver failure condition would just artificially inflate the values of sensitivity and specificity. For example, for such kind of validation sets, a predictive model can easily achieve 50% of sensitivity just by outputting '1' to the entire dataset without considering any input values/predictors. However, in the real world, a very poor prevalence of liver failure patients, between 1.0 to 5.0%, can be seen in the ICU patient population [53]. In this study, there were only a total of 755 patients diagnosed with liver failure in the ICU out of 81,135 patient admissions. The ratio (approx. 0.01) of patients with liver failure to those that do not have the condition is thus much less than the near 0.5 ratios implemented in most prior approaches. Hence, our study aims to address the previous limitation by validating the LFRI models with a representative real-world ICU patient dataset as described below.

Moreover, all of the prior approaches [18, 19, 20, 21, 22, 23] have considered only laboratory results as input features to estimate the likelihood of liver failure before its onset. Models developed during this effort have incorporated a more comprehensive set of model input features. The developed models provide a more comprehensive assessment of patient status by evaluating 24 patient data sources (i.e., model inputs) that include a range of laboratory results and vital signs. While increasing the number of required data sources by the model may be viewed as a potential limitation of the approach, it can also be viewed as a positive advancement. The incorporation of more model inputs provides the models with an improved ability to establish complex relationships between clinical factors that impact liver failure which is unaccounted for by some of the less sophisticated prior approaches developed and investigated to date.

The performance of the LFRI models developed during this effort can be improved by a number of methods in the future. Future efforts will investigate the further expansion of the set of model inputs used by the model. Some of the model inputs that can be considered in future efforts include but are not limited to: results from ALP, GGT, and lactate dehydrogenase (LD) tests, as well as globulin parameters, etc. The main challenge in expanding the feature set is to make sure that the data of these new features would be available across multiple datasets so that the generalizability of the developed models would not be compromised.

Further as mentioned in the section 'Introduction', apart from liver diseases, abnormal LFTs are indicative of many other diseases such as congestive heart failure, metastatic malignancy, etc. For example, Batin et al. completed an investigation and proved the prognostic importance of abnormal liver function tests, particularly AST and bilirubin, in *chronic heart failure* [54]. In general, the prognosis of patients with chronic heart failure is poor [55]. Although there are certain variables, such as left ventricular dysfunction, that can be used to predict the outcome obtaining the val-

ues for these variables involve specialized techniques and may not be available to the majority of patients. Hence, these simple LFTs which provide predictive information would be of obvious value. As the LFRI model developed in this approach considers these simple LFTs along with vital signs, this model could be useful to identify chronic heart failure patients by suspecting the incorrectly classified liver failure patients for this disease and confirming it further with the specialized techniques.

An ideal diagnostic model has high sensitivity combined with high specificity. However, models that are used in daily routine rarely conform to this criteria. Therefore, it is often necessary to find a sensible *trade-off* between the sensitivity and specificity to choose an appropriate model [56]. This implies that a model with high sensitivity can be achieved by compromising specificity and, conversely, a higher specificity can be achieved by accepting a lower sensitivity. When specificity which is related to the false positive rate (FPR = 1-Specificity) is low, it implicates a high frequency of false alarms that a health care provider has to respond, and this causes them *alarm fatigue*. Care providers with alarm fatigue tend to ignore or difficult to distinguish between alarms. This can result in a delay in intervention and patient harm, the US Food and Drug Administration specified a report which shows that there were 566 alarm-related deaths between 2005 and 2008 [57]. Hence, the extent to which the health care providers respond to an alarm, triggered by this model, related to a patient's high risk of developing liver failure is to be considered. At present, this is unknown and can be included in the factors which are independent of the prediction risk. Other such factors include frequency of alarms that a model should be generated to health care providers, the kind of mechanism used to convey these warnings, and the minimization of false alarms. Also, false alarms can be minimized by accepting a lower sensitivity model. For instance, in this approach, the MLP model has achieved a specificity of 77.5% which implies an FPR of 22.5%. This shows that 22.5% of alerts generated by this model are false alarms and this can be re-

duced by accepting a lower sensitivity model developed using GFF architecture (from Table 4.3, sensitivity=76.1% and specificity=81.6%) in this study. Investigation of all these issues and the work environment where the model needs to be used should be addressed before its deployment to avoid alarm fatigue and achieve an appropriate model in a clinical setting [58].

4.2.2 Limitations of the Effort

There are some limitations in our study. First, the prediction risk for each patient was validated only to evaluate the detection performance of the model. A potential study is needed to find whether and how this LFRI can impact remedial judgments. Second, the sensitivity and specificity of ICD-9 codes are diagnosis dependent [59]. In addition, these coding practices were prejudiced to the more frequently code, the more critical cases [60]. This limitation can be fixed by using automated tools to retrieve diagnosis-related information from the discharge notes [61].

Chapter 5

Conclusion and Future Works

In conclusion, neural network models developed in this effort have been demonstrated to predict the likelihood of liver failure for a patient in ICU many hours before standard screening protocols. This was accomplished by considering a number of data sources from the patient's electronic medical record (EMR) including but not limited to: laboratory results and vital signs. The performance of this model has also been validated externally using data of critical care patients from a completely different database and achieved a high sensitivity of 83.3% at a specificity of 77.5%. Moreover, this model has identified 83.5% (N=525) of liver failure patients with a median of 17.5 hours before the onset of liver failure. Achieving such a high performance when validated with an external database patient records substantiates that our approach has built a promising generalized model for predicting the liver failure in ICU population. Coordinating both evidence-based remedies and performance enhancement measures with such models can result in significant improvement of the outcome in ICU patients and helps in reaching the goal of learning health care systems.

Future research will be dedicated on the optimization of predictive accuracy of the model-generated LFRI. Model performance will be optimized in the future by investigating different sets of model inputs and predictors as described previously.

Another focus area of the future research will be on the development of patient

population-specific models. These types of models are gaining more popularity in research groups worldwide because of its potential to optimize the clinical treatment by predicting the outcomes of therapies, improving diagnosis, and informing the design of surgical platforms [62]. These models can be developed using training data of similar characteristics like morbidities, reasons for admission, age, gender, etc. For example, patient-specific models developed for alcoholic liver failure patients would be more accurate than the general model in predicting the risk of liver failure in alcoholic patients.

In addition, a limitation of this investigation was the ICD-9 coding practices. These codes were prejudiced to the more frequently codes which leads to more critical cases. This limitation will be addressed in future studies by considering automated tools to extract the diagnosis-related information from discharge notes.

While the models developed during this effort were designed for a very specific functionality (i.e., detection of risk for and the likelihood of liver failure), the LFRI modeling approach will be part of a much larger vision. Similar models predicting risk for organ failure or other patient outcomes will be integrated into a comprehensive clinical decision support system and patient monitoring tool for the ICU which will serve to identify at-risk patients and to focus treatment priorities for healthcare professionals. Efforts to develop such a comprehensive system are ongoing at the University of Toledo in close collaboration between the College of Engineering and the College of Medicine and Life Sciences.

References

[1] William Bernal, Georg Auzinger, Anil Dhawan, and Julia Wendon. "Acute liver failure". In: *The Lancet* 376.9736 (2010), pp. 190–201.

[2] Douglas G Farmer, Dean M Anselmo, Mark R Ghobrial, Hasan Yersiz, Suzanne V McDiarmid, Carlos Cao, Michael Weaver, Jesus Figueroa, Khurram Khan, Jorge Vargas, and others. "Liver transplantation for fulminant hepatic failure: experience with more than 200 patients over a 17-year period". In: *Annals of surgery* 237.5 (2003), p. 666.

[3] Anil C Anand, Peter Nightingale, and James M Neuberger. "Early indicators of prognosis in fulmitant hepatic failure: an assessment of the King's criteria". In: *Journal of hepatology* 26.1 (1997), pp. 62–68.

[4] HT Sørensen, JF Møller-Petersen, P Felding, C Andreasen, and JO Nielsen. "Epidemiology of abnormal liver function tests in general practice in a defined population in Denmark." In: *Danish medical bulletin* 38.5 (1991), pp. 420–422.

[5] Paul Sherwood, Iain Lyburn, Sandy Brown, and Stephen Ryder. "How are abnormal results for liver function tests dealt with in primary care? Audit of yield and impact". In: *BMJ: British Medical Journal* 322.7281 (2001), p. 276.

[6] Fredric D Gordon. *100 Questions & Answers About Liver Transplantation: A Lahey Clinic Guide*. Jones & Bartlett Learning, 2006.

[7] Miriam J Alter and Eric E Mast. "The epidemiology of viral hepatitis in the United States." In: *Gastroenterology Clinics of North America* 23.3 (1994), pp. 437–455.

[8] Paul Angulo. "Nonalcoholic fatty liver disease". In: *New England Journal of Medicine* 346.16 (2002), pp. 1221–1231.

[9] David A Rudnick and David H Perlmutter. "Alpha-1-antitrypsin deficiency: a new paradigm for hepatocellular carcinoma in genetic liver disease". In: *Hepatology* 42.3 (2005), pp. 514–521.

[10] Kevin Walsh and Graeme Alexander. "Alcoholic liver disease". In: *Postgraduate medical journal* 76.895 (2000), pp. 280–286.

[11] Flavio Lopes Ferreira, Daliana Peres Bota, Annette Bross, Christian Mélot, and Jean-Louis Vincent. "Serial evaluation of the SOFA score to predict outcome in critically ill patients". In: *Jama* 286.14 (2001), pp. 1754–1758.

[12] Hannah Lee, Susie Yoon, Seung-Young Oh, Jungho Shin, Jeongsoo Kim, Chul-Woo Jung, and Ho Geol Ryu. "Comparison of APACHE IV with APACHE II, SAPS 3, MELD, MELD-Na, and CTP scores in predicting mortality after liver transplantation". In: *Scientific Reports* 7.1 (2017), p. 10884.

[13] Evangelos B Cholongitas, Alex Betrossian, Gioacchino Leandro, Steve Shaw, David Patch, and Andrew K Burroughs. "King's criteria, APACHE II, and SOFA scores in acute liver failure". In: *Hepatology* 43.4 (2006), pp. 881–881.

[14] Ajay Duseja, Narendra S Choudhary, Sachin Gupta, Radha Krishan Dhiman, and Yogesh Chawla. "APACHE II score is superior to SOFA, CTP and MELD in predicting the short-term mortality in patients with acute-on-chronic liver failure (ACLF)". In: *Journal of digestive diseases* 14.9 (2013), pp. 484–490.

[15] Rahimi Dehkordi, Nasiri-Tousi Nasim, Azmoudeh Mohsen and Farid Ardalan. "Model for End stage Liver Disease (MELD) and Child-Turcotte-Pugh (CTP) scores: Ability to predict mortality and removal from liver transplantation waiting list due to poor medical conditions". In: *Archives of Iranian medicine* 17.2 (2014), p. 118.

[16] Russell Wiesner, Erick Edwards, Richard Freeman, Ann Harper, Ray Kim, Patrick Kamath, Walter Kremers, John Lake, Todd Howard, Robert M Merion, and others. "Model for end-stage liver disease (MELD) and allocation of donor livers". In: *Gastroenterology* 124.1 (2003), pp. 91–96.

[17] Scott W Biggins, W Ray Kim, Norah A Terrault, Sammy Saab,Vijay Balan, Thomas Schiano, Joanne Benson, Terry Therneau, Walter Kremers, Russell Wiesner, and others. "Evidence-based incorporation of serum sodium concentration into MELD". In: *Gastroenterology* 130.6 (2006), pp. 1652–1660.

[18] Ming Hua Zheng, Ke Qing Shi, Yu Chen Fan, Hai Li, Chao Ye, Qiong Qiu Chen, and Yong Ping Chen. "A model to determine 3-month mortality risk in patients with acute-on-chronic hepatitis B liver failure". In: *Clinical Gastroenterology and Hepatology* 9.4 (2011), pp. 351–356.

[19] Yi Ren, Lulu Liu, Ying Li, Fangwan Yang, Yihuai He, Yanping Zhu, Xinxin Hu, and Shide Lin. "Development and validation of a scoring system to predict progression to acute-on-chronic liver failure in patients with acute exacerbation of chronic hepatitis B". In: *Hepatology Research* (2018).

[20] Q-F Sun, J-G Ding, D-Z Xu, Y-P Chen, L Hong, Z-Y Ye, M-H Zheng, R-Q Fu, J-G Wu, Q-W Du, and others. "Prediction of the prognosis of patients with acute-on-chronic hepatitis B liver failure using the model for end-stage liver disease scoring system and a novel logistic regression model". In: *Journal of viral hepatitis* 16.7 (2009), pp. 464–470.

[21] Sheng-Jie Wu, Hua-Dong Yan, Zai-Xing Zheng, Ke-Qing Shi, Fa-Ling Wu, Yao-Yao Xie, Yu-Chen Fan, Bo-Zhi Ye, Wei-Jian Huang, Yong-Ping Chen, and others. "Establishment and validation of ALPH-Q score to predict mortality risk in patients with acute-on-chronic hepatitis B liver failure: a prospective cohort study". In: *Medicine* 94.2 (2015).

[22] J Rajanayagam, Eibe Frank, RW Shepherd, and PJ Lewindon. "Artificial neural network is highly predictive of outcome in paediatric acute liver failure". In: *Pediatric transplantation* 17.6 (2013), pp. 535–542.

[23] Moloud Abdar, Neil Yuwen Yen, and Jason Chi-Shun Hung. "Improving the Diagnosis of Liver Disease Using Multilayer Perceptron Neural Network and Boosted Decision Trees". In: *Journal of Medical and Biological Engineering* (2017), pp. 1–13.

[24] William Bernal, Catherine Hall, Constantine J Karvellas, Georg Auzinger, Elizabeth Sizer, and Julia Wendon. "Arterial ammonia and clinical risk factors for encephalopathy and intracranial hypertension in acute liver failure". In: *Hepatology* 46.6 (2007), pp. 1844–1852.

[25] Glòria Fernández-Esparrach, Alberto Sánchez-Fueyo, Pere Ginès, Juan Uriz, Llorenç Quintó, Pere-Joan Ventura, Andrés Cárdenas, Mónica Guevara, Pau Sort, Wladimiro Jiménez, and others. "A prognostic model for predicting survival in cirrhosis with ascites". In: *Journal of hepatology* 34.1 (2001), pp. 46–52.

[26] En-Qiang Chen, Fan Zeng, Ling-Yun Zhou, and Hong Tang. "Early warning and clinical outcome prediction of acute-on-chronic hepatitis B liver failure". In: *World journal of gastroenterology* 21.42 (2015), p. 11964.

[27] Ulrik Becker, Allan Deis, TI Sorensen, Morten Gronbaek, Knut Borch-Johnsen, Cecilia Florvall Muller, Peter Schnohr, and Gorm Jensen. "Prediction of risk of liver disease by alcohol intake, sex, and age: a prospective population study". In: *Hepatology* 23.5 (1996), pp. 1025–1029.

[28] Nathalie C Leite, Gil F Salles, Antonio LE Araujo, Cristiane A Villela-Nogueira, and Claudia RL Cardoso. "Prevalence and associated factors of non-alcoholic

fatty liver disease in patients with type-2 diabetes mellitus". In: *Liver International* 29.1 (2009), pp. 113–119.

[29] Bodil Andersson, Roland Andersson, Mattias Ohlsson, and Johan Nilsson. "Prediction of severe acute pancreatitis at admission to hospital using artificial neural networks". In: *Pancreatology* 11.3 (2011), pp. 328–335.

[30] Wan-dong Hong, Xiang-rong Chen, Shu-qing Jin, Qing-ke Huang, Qi-huai Zhu, and Jing-ye Pan. "Use of an artificial neural network to predict persistent organ failure in patients with acute pancreatitis". In: *Clinics* 68.1 (2013), pp. 27–31.

[31] Chirag R Parikh and Heather Thiessen-Philbrook. "Key concepts and limitations of statistical methods for evaluating biomarkers of kidney disease". In: *Journal of the American Society of Nephrology* 25.8 (2014), pp. 1621–1629.

[32] Jack V Tu. "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes". In: *Journal of clinical epidemiology* 49.11 (1996), pp. 1225–1231.

[33] Sonali B Maind, Priyanka Wankar, and others. "Research paper on basic of artificial neural network". In: *International Journal on Recent and Innovation Trends in Computing and Communication* 2.1 (2014), pp. 96–100.

[34] Rudolf Kruse, Christian Borgelt, Frank Klawonn, Christian Moewes, Matthias Steinbrecher, and Pascal Held. "Multi-layer perceptrons". In: *Computational Intelligence*. Springer, 2013, pp. 47–81.

[35] Matt W Gardner and SR Dorling. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences". In: *Atmospheric environment* 32.14-15 (1998), pp. 2627–2636.

[36] Andreas Zell. *Simulation Neuronaler Netze (1994)*.

[37] Ganesh Arulampalam and Abdesselam Bouzerdoum. "A generalized feedforward neural network classifier". In: *Neural Networks, 2003. Proceedings of the International Joint Conference on*. Vol. 2. IEEE. 2003, pp. 1429–1434.

[38] Xinghuo Yu, M Onder Efe, and Okyay Kaynak. "A general backpropagation algorithm for feedforward neural networks learning". In: *IEEE Transactions on Neural Networks* 13.1 (2002), pp. 251–254.

[39] Margaret H Dunham and Data Ming. *Introductory and advanced topics*. 2003.

[40] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3 (2016), p. 160035.

[41] Carole L Hart, David S Morrison, G David Batty, Richard J Mitchell, and George Davey Smith. "Effect of body mass index and alcohol consumption on liver disease: analysis of data from two prospective cohort studies". In: *Bmj* 340 (2010), p. c1240.

[42] Michael McShea, Randy Holl, Omar Badawi, Richard R Riker, and Eric Silfen. "The eICU research institute-a collaboration between industry, health-care providers, and academia". In: *IEEE Engineering in Medicine and Biology Magazine* 29.2 (2010), pp. 18–25.

[43] Ananth Ranganathan. "The levenberg-marquardt algorithm". In: *Tutoral on LM algorithm* 11.1 (2004), pp. 101–110.

[44] Luca Barnabei, Stefania Marazıa, and Raffaele De Caterina. "Receiver operating characteristic (ROC) curves and the definition of threshold levels to diagnose coronary artery disease on electrocardiographic stress testing. Part I: The use of ROC curves in diagnostic medicine and electrocardiographic markers of ischaemia". In: *Journal of Cardiovascular Medicine* 8.11 (2007), pp. 873–881.

[45] Jayawant N Mandrekar. "Receiver operating characteristic curve in diagnostic test assessment". In: *Journal of Thoracic Oncology* 5.9 (2010), pp. 1315–1316.

[46] Adelchi Azzalini. "The skew-normal distribution and related multivariate families". In: *Scandinavian Journal of Statistics* 32.2 (2005), pp. 159–188.

[47] Tony Badrick and Peter Turner. "Review and recommendations for the component tests in the liver function test profile". In: *Indian Journal of Clinical Biochemistry* 31.1 (2016), pp. 21–29.

[48] Eric Kipnis, Davinder Ramsingh, Maneesh Bhargava, Erhan Dincer, Maxime Cannesson, Alain Broccard, Benoit Vallet, Karim Bendjelid, and Ronan Thibault. "Monitoring in the intensive care". In: *Critical care research and practice* 2012 (2012).

[49] Barbara Hayes-Roth, Richard Washington, David Ash, Rattikorn Hewett, Anne Collinot, Angel Vina, and Adam Seiver. "Guardian: A prototype intelligent agent for intensive-care monitoring". In: *Artificial Intelligence in Medicine* 4.2 (1992), pp. 165–185.

[50] Blessing NR Jaja, Michael D Cusimano, Nima Etminan, Daniel Hanggi, David Hasan, Don Ilodigwe, Hector Lantigua, Peter Le Roux, Benjamin Lo, Ada Louffat-Olivares, and others. "Clinical prediction models for aneurysmal subarachnoid hemorrhage: a systematic review". In: *Neurocritical care* 18.1 (2013), pp. 143–153.

[51] Gary S Collins, Omar Omar, Milensu Shanyinde, and Ly-Mee Yu. "A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods". In: *Journal of clinical epidemiology* 66.3 (2013), pp. 268–277.

[52] SE Bleeker, HA Moll, EW Steyerberg, ART Donders, Gerarda Derksen-Lubsen, DE Grobbee, and KGM Moons. "External validation is necessary in prediction research:: A clinical example". In: *Journal of clinical epidemiology* 56.9 (2003), pp. 826–832.

[53] Lior Fuchs, Catherine E Chronaki, Shinhyuk Park, Victor Novack, Yael Baumfeld, Daniel Scott, Stuart McLennan, Daniel Talmor, and Leo Celi. "ICU admission characteristics and mortality rates among elderly and very elderly patients". In: *Intensive care medicine* 38.10 (2012), pp. 1654–1661.

[54] P Batin, M Wickens, D McEntegart, L Fullwood, and AJ Cowley. "The importance of abnormalities of liver function tests in predicting mortality in chronic heart failure". In: *European heart journal* 16.11 (1995), pp. 1613–1618.

[55] SOLVD Investigators*. "Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure". In: *New England Journal of Medicine* 325.5 (1991), pp. 293–302.

[56] Christopher M Florkowski. "Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests". In: *The Clinical Biochemist Reviews* 29.Suppl 1 (2008), S83.

[57] Kierra Jones. "Alarm fatigue a top patient safety hazard". In: *CMAJ: Canadian Medical Association Journal* 186.3 (2014), p. 178.

[58] Heleen Van Der Sijs, Jos Aarts, Arnold Vulto, and Marc Berg. "Overriding of drug safety alerts in computerized physician order entry". In: *Journal of the American Medical Informatics Association* 13.2 (2006), pp. 138–147.

[59] James R Campbell and TH Payne. "A comparison of four schemes for codification of problem lists." In: *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association. 1994, p. 201.

[60] Stacey-Ann Whittaker, Mark E Mikkelsen, David F Gaieski, Sherine Koshy, Craig Kean, and Barry D Fuchs. "Severe sepsis cohorts derived from claims-based strategies appear to be biased towards a more severely ill patient population". In: *Critical care medicine* 41.4 (2013).

[61] Suchi Saria, Gayle McElvain, Anand K Rajani, Anna A Penn, and Daphne L Koller. "Combining structured and free-text data for automatic coding of patient outcomes". In: *AMIA Annual Symposium Proceedings*. Vol. 2010. American Medical Informatics Association. 2010, p. 712.

[62] Maxwell Lewis Neal and Roy Kerckhoffs. "Current progress in patient-specific modeling". In: *Briefings in bioinformatics* 11.1 (2009), pp. 111–126.