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 preiii 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.

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 1.1 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 2 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 3 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]. 4 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. 5 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-TurcottePugh (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 6 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 7 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 8 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 hybridiza9 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. 10 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.

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 15 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 wij 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 (xi) is processed by j th node in the next layer and Figure 3-2 represents this process.

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]. 19 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. 20 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 21 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 MAT LAB R (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 openaccess 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 vali22 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 23 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 R 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]. 24 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 (φLF RI ) 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 (φLF RI ). 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 summa25 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.