RESEARCH ARTICLE

An artificial neural network approach to diagnose and predict liver dysfunction and failure in the critical care setting

Pappada S[1-3], Sathelly B[3], Schmieder J[4], Javaid A[3], Owais M[3], Cameron B[2], Khuder S[5], Kostopanagiotou G[6], Smith R[7,8], Sparkle T[1], Papadimos T[1,6,9]

[1]Department of Anesthesiology, College of Medicine and Life Sciences
[2]Department of Bioengineering
[3]Department of Electrical Engineering and Computer Science
[4]College of Medicine and Life Sciences
[5]Department of Medicine, College of Medicine and Life Sciences
University of Toledo, Toledo, Ohio, USA
[6]2nd Department of Anesthesiology, National and Kapodistrian University of Athens, Attikon University Hospital, Athens, Greece
[7]Department of Psychiatry
[8]Department of Neurosciences
[9]Department of Surgery
University of Toledo, Toledo, Ohio, USA

Abstract

Background: Detecting liver dysfunction/failure in the intensive care unit poses a challenge as individuals afflicted with these conditions often appear symptom-free, thereby complicating early diagnoses and contributing to unfavorable patient outcomes. The objective of this endeavor was to improve the chances of early diagnosis of liver dysfunction/failure by creating a predictive model for the critical care setting. This model has been designed to produce an index that reflects the probability of severe liver dysfunction/failure for patients in intensive care units, utilizing machine learning techniques.

Materials and Methods: This effort used comprehensive open-access patient databases to build and validate machine learning-based models for predicting the likelihood of severe liver dysfunction/failure. Two artificial neural network model architectures that derived a novel 0-100 Liver Failure Risk Index were developed and validated using the comprehensive patient databases. Data used to train and develop the models included clinical (patient vital signs) and laboratory results related to liver function which included liver function test results. The performance of the developed models was compared in terms of sensitivity, specificity, and the mean lead time to diagnosis.

Results: The best model performance demonstrated an 83.3 % sensitivity and a specificity of 77.5 % in diagnosing severe liver dysfunction/failure. This model accurately identified these patients a median of 17.5 hours before their clinical diagnosis, as documented in their electronic health records. The predictive diagnostic capability of the developed models is crucial to the intensive care unit setting, where treatment and preventative interventions can be made to avoid severe liver dysfunction/failure.

Conclusion: Our machine learning approach facilitates early and timely intervention in the hepatic function of critically ill patients by their healthcare providers to prevent or minimize associated morbidity and mortality. HIPPOKRATIA 2024, 28 (1):1-10.

Keywords: Clinical decision support, intelligent diagnostics, machine learning, liver dysfunction, liver failure

Corresponding author: Thomas Papadimos, MD, MPH, FCCM, Professor, Departments of Anesthesiology and Surgery, 3000 Arlington Ave, Mail Stop 1137, Toledo, Ohio, USA 43614, e-mail: Thomas.Papadimos@utoledo.edu

Introduction

The liver plays a crucial role in both metabolic and detoxification processes. Its primary functions include production of bile, clotting factors, albumin, and the elimination of bilirubin[1]. Acute liver failure (ALF) is a condition defined as the rapid development of severe liver dysfunction. Usually, ALF is characterized by abnormal coagulation, which manifests as an elevated prothrombin time (PT) or international normalized ratio (INR), often accompanied by altered mentation[2-4]. This condition is associated with a very high mortality rate that is estimated to be as high as 90 % without liver transplantation[2], and can affect up to 3,000 patients annually in the United States[2,5,6]. Although liver failure is commonly categorized based on the interval between the time of developing the first symptom[7], the term ALF is used to refer to three types of the conditions: 1) hyperacute jaundice-to-encephalopathy interval of less than seven days, 2) acute, and 3) subacute[8,9]. Hyperacute ALF is commonly caused by hepatitis A, acetaminophen overdose, and ischemia, while the acute and subacute types are associated with hepatitis B, herbal therapies, and autoimmune hepatitis[7].

Early detection of hepatic failure in the intensive care unit (ICU) is challenging, even with appropriate laboratory tests. Abnormal liver function tests (LFTs) may indicate other diseases besides those related to the liver[10-13]. These LFTs sometimes can be misleading and may result in inappropriate treatments. Therefore, an earlier and more accurate detection of liver dysfunction can be critical. While lab testing may detect various liver diseases such as cirrhosis, viral hepatitis, alcoholic and non-alcoholic related liver diseases[14-17], most ALF diagnostic work has been limited to detecting patients that already have some aspect of liver failure.

Many systems have been used to assist in the evaluation of the illness severity among liver failure patients[18,19]. However, these scoring systems limit their usefulness to the evaluation of severity and morbidity, and not to identifying the risk of acute liver conditions. Although early warning systems have been developed to improve the early detection of patients with liver failure, they were developed only for liver failure associated with particular conditions or disease states[20-23]. More specifically, scoring systems have been developed to aid in the diagnostic approaches, prognostication, and transplantation for those patients with non-alcoholic fatty liver disease and cirrhosis[23,24]. Complex physiological systems and their parameters can be modeled using machine learning approaches such as artificial neural networks (ANN). These types of models are well suited for the diagnosis and prediction of various disorders because they consider the effect and relationships between variables and parameters that may not be as significant when compared to conventional statistical methods; ANNs have been shown to be even more effective than multivariate logistic regression (LR) models of disease[25]. Several reports exist regarding the development of ANN-based models for diagnosing serious states of liver disease[15,17,26]. These earlier works neither assist with the diagnosis nor are they predictive of liver dysfunction/failure amongst an ICU patient population, but rather support the diagnosis of liver disease broadly or predict poor outcomes.

There is no well-defined diagnostic tool for severe liver dysfunction/failure in an ICU[27]. However, improved medical treatments of patients with ALF has resulted in better patient outcomes[28]. Providers rely on a combination of laboratory evaluations and LFT in addition to reviewing scoring systems related to organ function and patient mortality. Such systems include the sequential organ failure assessment (SOFA) score, model for end-stage liver disease (MELD) score, and the acute physiology and chronic health evaluation (APACHE) score. These scoring systems are usually calculated infrequently (e.g., every 24 hours) and are generally linked to patient mortality rather than diagnosis of liver dysfunction/failure[16,18,19]. To our understanding, there is no intuitive clinical index or marker related directly to the likelihood of a patient acquiring severe liver dysfunction/failure. In this study we attempt to address this deficiency in the clinical arena by generating a liver failure risk index (LFRI) derived via a machine learning model (ANN). The LFRI is calculated hourly using 29 up-to-date patient medical records data streams and is intended as a temporal intuitive marker representing the likelihood or probability that an individual patient will experience severe liver dysfunction/failure. The intent of this LFRI is to provide healthcare providers (HCPs) in the critical care setting with a clinically relevant indicator of patient liver status. In this study, we also explore the potential of the LFRI in providing predictive capacity (i.e., predictive diagnosis) of severe liver dysfunction/failure prior to a clinical diagnosis being documented in the patient's medical record. Currently, no tools provide prognostication of liver dysfunction/failure in the critical care setting, and our approach in using the LFRI intends to address this need. Predicting the onset of liver dysfunction/failure would allow HCPs to make preventative and treatment decisions/interventions to improve patient safety and outcomes[28].

Methods

Model training and validation set development

To develop and validate the machine learning-based models (Figure 1) that were required to generate the proposed LFRI, we leveraged two large open-access critical care databases: i) the Medical Information Mart for Intensive Care III (MIMIC-III)[29], and ii) the eICU Collaborative Research Database[30]. We used the former for model training, and the latter to validate the developed models.

MIMIC-III consists of more than 45,000 critical care patients admitted from the ICUs of the Beth Israel Deaconess Medical Center in Boston, MA between 2001-2012 and was used to train the models[29]. We generated the model training sets for each ANN using a custom software application developed utilizing MATLAB (Mathworks, Natick, MA, USA). Model inputs included

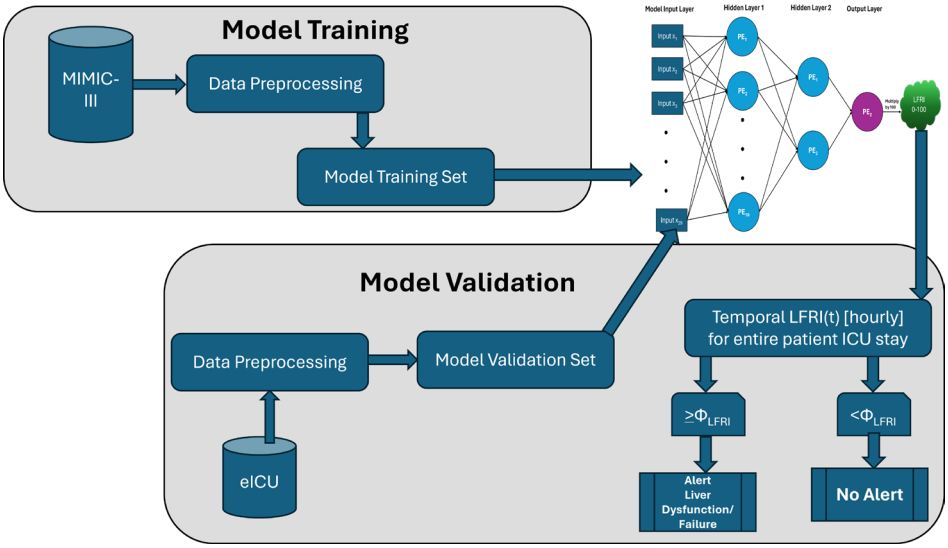

Figure 1: Overview of model development (training) and validation processes of the liver failure risk index model.
MIMIC-III: medical information mart for intensive care III, LFRI: liver failure risk index, Φ_{LFRI}: optimal threshold value for the liver failure risk index.

bedside patient vital signs and their corresponding three-hour intervals (current vital sign at time *t* and vital sign values at *t*-1 and *t*-2 hours) collected during a patient's ICU length of stay. Vital signs used for the models included heart rate, systolic blood pressure, diastolic blood pressure, temperature, respiratory rate, and blood oxygen saturation level (SpO_2). In addition to vital signs, the patient's laboratory and liver function test (LFT) results were also used. Laboratory and LFT results used included: albumin, PT, partial thromboplastin time (PTT), alanine aminotransferase (ALT), INR, total bilirubin, direct bilirubin, aspartate aminotransferase (AST), serum creatinine, sodium, and blood urea nitrogen (BUN). The 29 model input features were defined not by a formal feature selection process but were based on the Delphi method[30] (a structured communication that is a systematic, interactive method that relies on a panel of subject matter experts - which consisted of the physician members of our team). These features were chosen based on their relationship with liver function and overall patient status (indicated by vital signs available hourly in each of the datasets used).

To develop the required dataset for training ANN-based models that output the targeted LFRI, we evaluated up-to-date patient data hourly concerning liver failure diagnostic criteria provided by collaborating physicians. Hourly evaluation of patient data was completed via a MATLAB script which cycled through all patient records, hour by hour, in the MIMIC-III dataset. Included in the MIMIC-III dataset were ICD-9 diagnosis codes 570-573 used for liver dysfunction/failure conditions. If patients during their ICU stay had an ICD-9 diagnosis and met the defined clinical diagnostic criteria at a timestamp, a "1" was used as the target model output at each timestamp where diagnostic criteria were met. Diagnostic criteria for liver failure were defined and agreed upon *a priori* given abnormal laboratory and LFT results. If two or more of the laboratory/test results were out of range at any timestamp, the patient met diagnostic criteria. Table 1 includes laboratory and LFT results, which were evaluated hourly along with their corresponding upper and lower limits, which defined normal ranges. If the patient exceeded these normal ranges, they met diagnostic criteria for a particular laboratory result. If two or more laboratory results were abnormal, and the patient

Table 1: Patient laboratory results and tests were used to evaluate liver dysfunction/failure and generate model training sets. Units differ based on the laboratory or test result to which they apply and are only represented in whole-number units. For example, serum creatinine and BUN results are represented in mg/dL.

Feature	Lower Limit	Upper Limit
Prothrombin Time	11	13.5
PTT	25	35
INR	0.8	1.1
Total Bilirubin	0.3	1.9
AST	10	34
ALT	7	56
Serum Creatinine	0.7	1.3
Sodium	136	145
Albumin	3.4	5.4
Direct Bilirubin	0	0.4
BUN	6	24

PTT: partial thromboplastin time, INR: international normalized ratio, AST: aspartate aminotransferase, ALT: alanine transaminase, BUN: blood urea nitrogen.

had a liver dysfunction/failure diagnosis documented in their medical record (via ICD-9 codes defined previously), the target or desired model output was held at "1" throughout the remaining ICU stay. A "0" was used as the target or desired model output when the patient did not have a documented diagnosis and model diagnostic criteria were not met. No patient datasets were removed from the MIMIC-III dataset used for model training as no time-stamped diagnosis codes (ICD) were present in the dataset. Thus, we used the above approach to generate a uniform training set based on the provided diagnostic criteria.

The second database used was the eICU Collaborative Research Database[31]. Patients were excluded from the validation set if they had a preexisting diagnosis of liver dysfunction before admission or were assigned a diagnosis of liver dysfunction less than four hours into their ICU admission (this was because of the predictive intent of this model). Models were validated using a total of 81,135 patients from this dataset, which was the number of patients that resulted from excluding patients with a diagnosis of liver dysfunction/failure within the first four hours of ICU admission. This resulted in a total of 755 patients who acquired liver dysfunction or failure during their ICU admission.

Patients were excluded from the validation dataset if they had a diagnosis of liver failure less than four hours into their ICU admission (i.e., considered a preexisting diagnosis) which was either documented in the patient record or based on the diagnostic criteria (that included hourly evaluation of patient data) discussed previously. As such, the validation dataset used is representative of a broad population of critically ill patients [e.g., postoperative\surgical, cardiac\cardiovascular, trauma, medical, and general (mixed medical/surgical)] with various reasons for admission. Specific patient characteristics of the overall patient populations are included in publications (and accompanying websites) cited for each respective dataset referenced in this manuscript. For the purposes of this study, patient demographics and reasons for admission were not tracked or analyzed, and all available patient records were used to provide a universally applicable model for a broad and diverse ICU population. This supports a targeted evaluation of model performance across multiple institutions to better ensure the utility of the approach outside of a single institution. For the purposes of this study model, input features were defined based on subject matter expertise provided by the team of HCPs supporting the effort. Additionally, model inputs were chosen for inclusion as they represent standard patient data streams routinely collected in the ICU and available across any ICU. Both the MIMIC-III and eICU databases were used because they provided the chosen model input features hourly.

It is also worth noting that model training and validation sets were generated using the last known laboratory value, which created sufficiently large datasets to develop and validate the models described in this study using data collected every hour during each patient's ICU length of stay. Missing laboratory or data results were encoded with a value of "-1" such that the model output could still be generated in the absence of model input data. Using the last known laboratory results and encoding missing data was implemented, as shown in Figure 1.

Neural Network Model Development

In this study, we used a commercially available software application, NeuroSolutions® (NeuroDimension, Gainesville, FL, USA), to generate the neural network models. We developed a multilayer perceptron (MLP) and a generalized feed-forward (GFF) ANN. Models were trained with the Levenberg-Marquardt (LM) algorithm. All ANNs were configured with batch training and designed to terminate training if the mean squared error was less than 0.1 or after a total of 1,000 epochs, whichever came first. Models were developed to include two hidden layers with 10 and two processing elements in the first and second hidden layers of the model architecture, respectively.

All ANN models were designed to derive a 0-100 LFRI hourly based on previously detailed input features. The derivation of the 0-100 LFRI was enabled by the model training set process discussed previously, which provides hourly desired model outputs as either "0" (patient does not meet liver dysfunction/failure diagnostic criteria present at current hourly timestamp) or "1" (patient meets liver dysfunction/failure diagnostic criteria at hourly timestamp). Regarding ANN model architecture, a sigmoid transfer function was used for processing elements in the models' hidden and output layers that served to constrain or normalize model inputs and outputs between a range of zero to one. The model output (i.e., output activation of the ANN) was thus a value ranging between zero and one. Values closer to "0" indicate a lower probability of liver dysfunction/failure, and values closer to "1" indicate a higher probability or likelihood that the patient is in a state of liver dysfunction/failure or trending towards this state. As such, the LFRI was derived by multiplying the output of the ANN by 100 to achieve the LFRI on the intended 0-100 scale. Based on model training criteria, the output of the ANN is related to the probability or likelihood of class membership, i.e., "0" being no liver dysfunction/failure and "1" indicating the patient has liver dysfunction/failure. As model inputs are evaluated hourly, the model-derived LFRI provides the intended intuitive temporal index related to the patient liver function that can be used for diagnosis and predictive diagnosis of liver dysfunction/failure.

Analysis and Validation

The ANN model architectures investigated during this effort were separately subjected to performance analyses that included the generation of Receiver Operating Characteristic (ROC) curves and the determination of an optimal threshold value for the LFRI (Φ_{LFRI}) for use in alerting HCPs of current or impending liver dysfunction/failure. Factors such as LFRI sensitivity, specificity to

the false alarm rate, and mean and median lead times to clinical diagnosis of liver failure were used to determine an optimal Φ_{LFRI} for use in model-assisted diagnosis. The area under the ROC curve (AUC) was calculated to evaluate the overall diagnostic capabilities of the LFRI. Another key model performance metric was predictive capacity, which refers to the percentage of patients correctly diagnosed with liver dysfunction/failure before a diagnosis was documented in the patient record (based on timestamped diagnoses present in the eICU dataset). This performance metric differs from sensitivity in that sensitivity refers to an accurate diagnosis at any time point, including after clinical diagnosis is reached. For this study, we used Φ_{LFRI} =50 as the optimal model output threshold for diagnosis. Values exceeding Φ_{LFRI} represent a liver dysfunction/failure diagnosis based on the model output. Model timestamps exceeding the defined Φ_{LFRI} were used to calculate lead times to diagnosis by subtracting the timestamp of the model-generated diagnosis (based on Φ_{LFRI}) from the timestamp of the diagnosis documented in the validation dataset (eICU). Negative values indicated that model-generated diagnosis was predictive of liver dysfunction/failure, whereas positive values indicated delayed but correct diagnoses. Of these predictive diagnoses, the median, mean, and standard deviation (SD) of the lead time to diagnosis was calculated for each model.

Results

Model performance analysis demonstrated little difference in performance between the two models using ROC analysis, and the MLP (AUC: 0.8622) narrowly outperformed the GFF model (AUC: 0.8618). Figure 2 includes plots of the ROC curves for both the MLP model (Figure 2 left) and the GFF ANN model (Figure 2 right).

As previously discussed, a Φ_{LFRI} value of 50 was chosen as it provided a balance of sensitivity and specificity with a false-positive rate (FPR) of <25 %, which was targeted for this effort. A 25 % FPR was chosen to mitigate the incidence of false alarms in patient care settings. The likelihood ratios for positive and negative test results for each model can be calculated using the sensitivity and specificity results presented in Table 2. The likelihood ratio for positive test results for each model was calculated as 3.69 and 4.16 for the MLP and GFF models, respectively. The likelihood ratio for negative test results for each model was calculated as 0.22 and 0.29 for the MLP and GFF models, respectively. Table 2 summarizes all model performance results based on the chosen Φ_{LFRI} of 50 for each ANN model developed in this effort.

Both models detected a high percentage of the 755 patients in the validation dataset who experienced liver failure during their ICU stay. Over 80 % of the LFRI-based alerts were enacted before a documented clinical diagnosis, demonstrating the predictive capacity of the LFRI. In the leading model, the mean lead time to diagnosis was calculated as 34.4 ± 27.7 hours, and due to the skewed distribution, the mean and median lead time to diagnosis differed by greater than sixteen hours. As such, the median lead time to diagnosis better represented an overall model performance. The median lead time to diagnosis for the MLP ANN was calculated as 17.5 hours.

The LFRI is intended to be an intuitive numerical index related to the likelihood of a patient experiencing liver dysfunction/failure. The LFRI intends for HCPs to track and evaluate this index over time and provide preventative or treatment decisions to preserve a patient's liver function. Figure 3, Figure 4, and Figure 5 demonstrate plots of the model-generated LFRI over time for some example patient cases. These figures are chosen as they represent each distinct model output performance evaluation criterion, including true positive predictive diagnoses, true positive delayed diagnoses, false negative

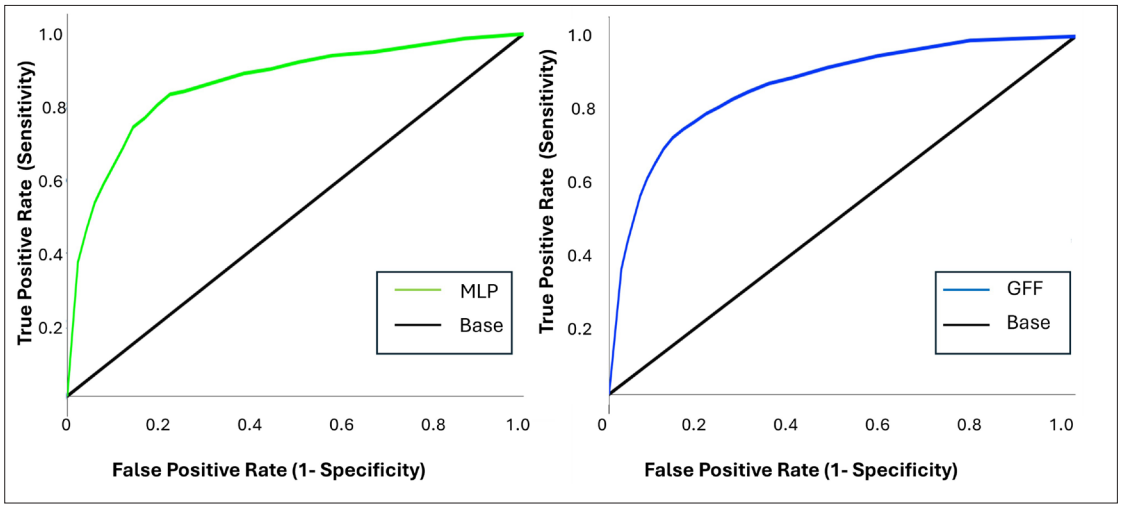

Figure 2: Receiver operating characteristic (ROC) curves and corresponding base (calibration) curves for developed multilayer perceptron (MLP) model and generalized feed-forward neural network (GFF) model. The area under the ROC curves (AUC) for the MLP model was 0.8622, and the AUC for the GFF model was 0.8618.

MLP: multilayer perceptron, GFF: generalized feed-forward.

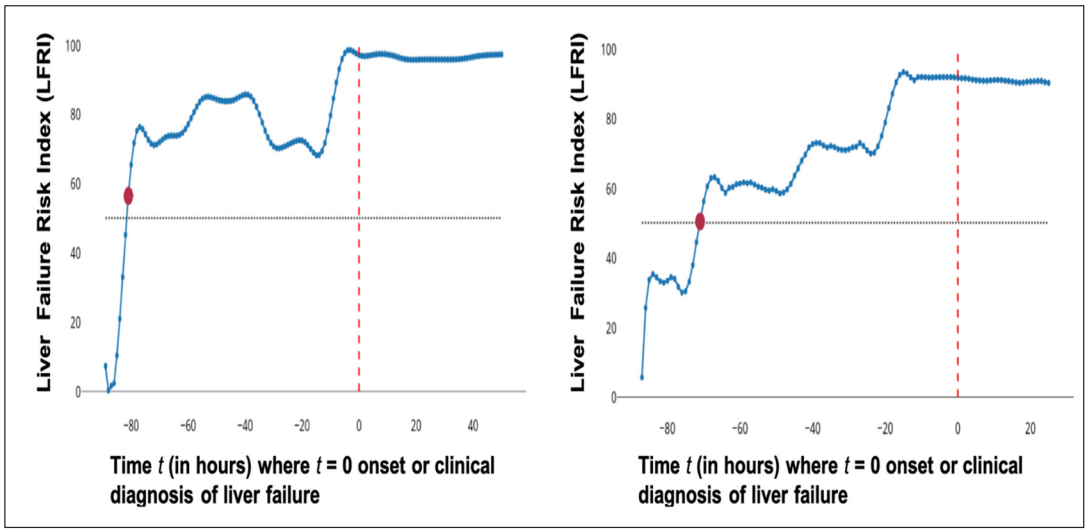

Figure 3: Graphs demonstrate two patient cases where the model-generated liver failure risk index (LFRI) predicted liver failure. Included are graphs of two patients who were diagnosed with liver failure during their intensive care unit length of stay. In these two patient cases, the LFRI was predictive of liver failure before its clinical diagnosis or onset. The horizontal dashed black line in these figures represents the optimal threshold value for the liver failure risk index (Φ_{LFRI}) value of 50. The circle, when present, indicates the time at which the model generated LFRI exceeded the defined Φ_{LFRI}, and a healthcare provider would have been initially alerted. The vertical dotted line in the figures represents the timestamp where the initial clinical diagnosis of liver failure was documented in the dataset.

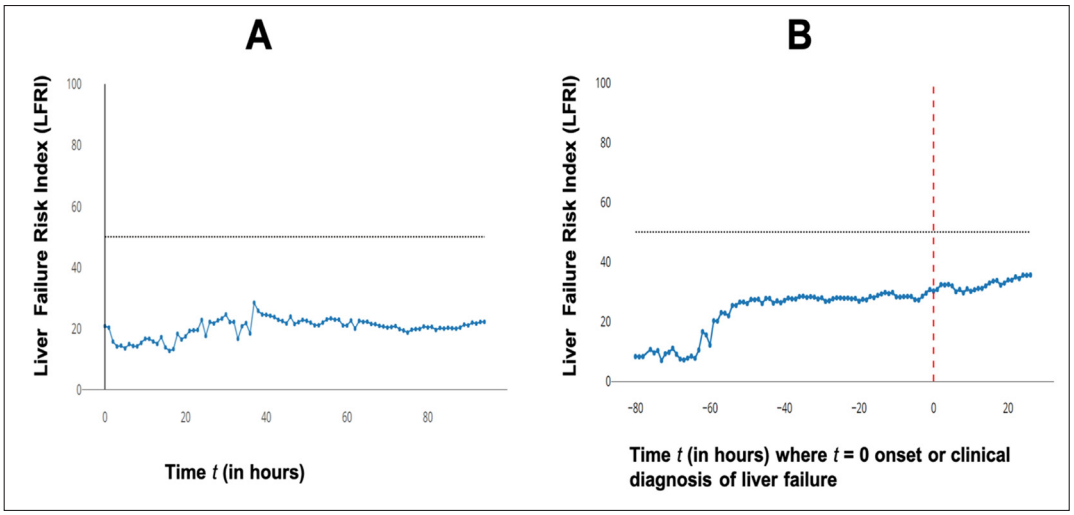

Figure 4: Model-generated liver failure risk index (LFRI) values for two distinct patient cases. Figure 4A demonstrates a true negative case in a patient with no liver dysfunction and no diagnosis of liver failure where the LFRI did not exceed the optimal threshold value for the liver failure risk index (Φ_{LFRI}). Figure 4B represents a false-negative case where LFRI did not alert providers of liver failure, despite a documented clinical diagnosis. The horizontal dashed black line in these figures represents the Φ_{LFRI} value of 50. The circle, when present, indicates the time at which the model generated LFRI exceeded the defined Φ_{LFRI}, and a healthcare provider would have been initially alerted. The vertical dotted line in the figures represents the timestamp where the initial clinical diagnosis of liver failure was documented in the dataset.

diagnoses (missed diagnoses), true negative diagnoses, and false positive diagnoses.

Discussion

To provide predictive or improved diagnostics for severe liver dysfunction/failure, this study evaluated the efficacy and effectiveness of machine learning-based approaches, specifically ANNs. Earlier approaches implemented various machine learning algorithms to develop predictive models for early identification of patient outcomes resulting from liver disease and liver failure; however, most models have not been tailored for use in an ICU. The models developed and validated in this initial study have shown potential for both the effective and predictive diagnosis of liver dysfunction/failure in a critical care patient population.

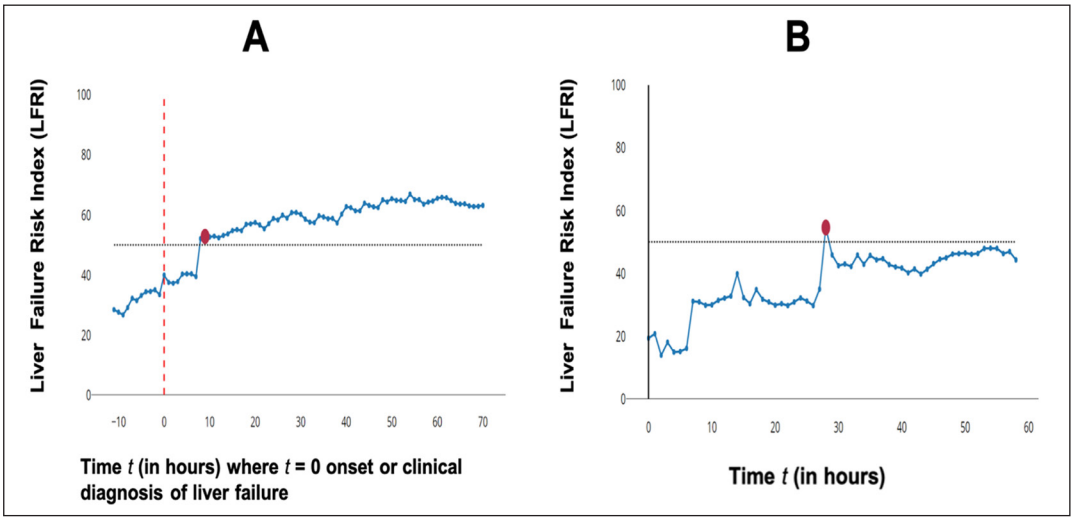

Figure 5: Model-generated liver failure risk index (LFRI) values for two distinct patient cases. Figure 5A demonstrates a true positive diagnosis (but delayed diagnosis) that occurred approximately ten hours after the documented clinical diagnoses and the onset of liver failure. This was characterized as a correct but non-predictive diagnosis. Figure 5B includes a false-positive diagnosis of liver failure by the LFRI, where the LFRI exceeded the threshold with no documented clinical diagnosis present in the patient record. In these figures, the horizontal dashed black line represents the optimal threshold value for the liver failure risk index (Φ_{LFRI}) value of 50. The circle, when present, indicates the time at which the model generated LFRI exceeded the defined Φ_{LFRI}, and a healthcare provider would have been initially alerted. The vertical dotted line in the figures represents the timestamp where the initial clinical diagnosis of liver failure was documented in the dataset.

Table 2: Liver failure risk index (LFRI) model performance measures, including sensitivity, specificity, and predictive capacity, derived from model validation via the eICU Collaborative Research Database across multiple intensive care units (ICUs).

Model	True Positives	False Positives	True Negatives	False Negatives	Sensitivity	Specificity	AUC	Likelihood Ratio (Positive Tests)	Likelihood Ratio (Negative Tests)	Predictive Capacity
MLP	629	18084	62296	126	83.3 %	77.5 %	0.8622	3.69	0.22	83.5 % (n =525)
GFF	575	14736	65644	180	76.1 %	81.7 %	0.8618	4.16	0.29	80.2 % (n =461)

MLP: multilayer perceptron, GFF: generalized feed-forward, AUC: area under the receiver operating characteristic (ROC) curve.

The most accurate ANN developed in this study was able to diagnose over 83 % of patients who experienced liver dysfunction/failure during their ICU stay. Greater than 82 % of correct diagnoses demonstrated predictive capacity, i.e., were predictive of liver dysfunction/failure. Although there was a false positive rate of 22.5 %, the models maintained a clinically applicable balance of sensitivity and specificity. It is worth noting that the model performance achieved during this effort is comparable to the performance of other machine learning-based models reported in the literature by Nemati et al who developed a machine learning-based model to predict onset of sepsis and achieved similar AUC[32].

As mentioned above, the severity of illness and patient mortality are currently evaluated by different scoring systems such as SOFA, APACHE, etc[18,19]. These scoring systems have little utility in exclusively predicting or diagnosing liver dysfunction/failure. Although early warning systems have been developed to improve the early detection of patients with liver failure[14,20-22] they were developed only for liver failure associated with particular conditions or disease states. More specifically, scoring systems have been developed to aid in the diagnostic approaches, prognostication, and transplantation for those patients with non-alcoholic fatty liver disease and cirrhosis[23,24]. Complex physiological systems and their parameters can be modeled using machine learning approaches such as an ANN. These types of models are well suited for the diagnosis and prediction of various disorders because they consider the effect and relationships between variables and parameters that may not be as significant compared to conventional statistical methods; ANNs have been shown to be even more effective than multivariate Logistic Regression (LR) models of disease[25]. Several reports exist regarding developing ANN-based models for diagnosing serious states of liver disease[15,19,26]. However, as previously highlighted, these studies were neither diagnostic of liver dysfunction/fail-

ure nor predictive of liver dysfunction/failure in an ICU patient population. They largely only supported a diagnosis of liver disease or predicted poor outcomes.

There is no well-defined diagnostic tool for severe liver dysfunction/failure in an ICU; however, recent medical treatment of patients with ALF has improved patient outcomes[28]. Providers rely on the evaluation of general laboratory and liver function tests and reviewing scoring systems related to organ function and patient mortality including SOFA, MELD, APACHE, and others. The MELD scoring system is explicitly used to assign priority to liver transplant candidates and is calculated via laboratory/test results of bilirubin, serum sodium, INR, albumin, and serum creatinine. Many of the test results used in the calculation of a MELD score were applied as input features for our developed model. A primary limitation of these scoring systems is their relatively infrequent calculation (e.g., every 24 hours), and they are related to longer-term patient outcomes such as mortality rather than short-term measures of organ status[16,18,19]. As such, these scoring systems do not thoroughly assess patient liver function. For instance, in the case of SOFA, only a single dimension is used to evaluate liver function (bilirubin). Assessing only a single value or a small subset of factors may not provide a complete picture required for diagnosis. The machine learning ANN-based models developed in our study anticipate patterns across multiple patient data sources to provide a more complete diagnostic picture.

The LFRI derived by the models generated in this study, provide an intuitive, clinically relevant marker of the likelihood that a patient has or will experience liver dysfunction/failure in the ICU.

While the use of artificial intelligence and machine learning approaches in healthcare has experienced tremendous growth recently, model validation methods should be carefully considered to ensure the real-world applicability of developed solutions[33]. Refraining from developing machine learning-based models using relatively small sample-sized retrospective datasets acquired from single centers for training and validation is imperative. To this end, our study accessed two open-access databases with significantly large patient populations for model training and validation. Models were trained using the MIMIC-III dataset, and model validation was supported using the eICU collaborative research database. This database included patient data collected across multiple ICUs located throughout the United States. Thus, our methodology tempers the potential for a model to be generalized to a particular institution or a specific patient population. This increases the likelihood that the developed models may be applicable and provide better clinical utility when implemented at other healthcare institutions.

It is also important to avoid testing a developed predictive model with a validation dataset that has an equal ratio of conditions. This can artificially inflate model sensitivity and specificity. For example, a predictive model can easily achieve 50 % sensitivity via a constant output of "1" for an entire dataset without considering any input values. In this study, <1 % of patients were diagnosed with liver dysfunction which is much less than the incidence and ratio of disease and non-disease patients used in prior approaches[14,15]. We eliminated patients from our model validation set who had a preexisting diagnosis of liver dysfunction/failure less than four hours into their ICU admission. As such, the 81,000+ patients used for model validation should be representative of the distribution of patients who acquire severe liver dysfunction/failure during their ICU admission.

Our study is not without limitations. This study only provides a retrospective analysis of model performance, and the accuracy depends on the existing data sources and diagnoses in the datasets. The capabilities of the models rely on the fact that the data was accurately and promptly recorded into the patient's electronic health record (EHR), which may or may not be the case. The models only used concurrent EHR data in terms of patient laboratory results and a three-hour history of patient vital signs collected via bedside monitoring. Performance can perhaps be improved in the future by incorporating a history of patient labs, as well as a potential expansion to other relevant patient data that was not identified by our team of collaborating HCPs for this initial study. Future efforts will investigate the further expansion of the set of model inputs used by the model, including employing feature selection methods on a more comprehensive patient EHR dataset. A primary challenge will be expanding the feature set to ensure that new features incorporated for use by the models are consistently collected and readily available across multiple institutions. In doing so, the generalizability of future developed models will not be compromised.

The simplistic alerting method implemented during this initial study and model development effort also has limitations. Improvements to the alerting mechanism modulated by values of the model-derived LFRI will be the subject of future investigation. This preliminary effort only investigated using a static value 50 for the Φ_{LFRI}. This means that once the threshold is reached, an alert will be provided to evaluate the patient for liver dysfunction. This can be improved in the future by providing an alert when several consecutive values exceed a threshold or when the threshold is exceeded for a defined time period. Choosing an appropriate model is often necessary to find a sensible trade-off between sensitivity and specificity. When specificity is low, it delivers a high number of false alarms, resulting in *alarm fatigue.* If alarms are disregarded by healthcare professionals' patient outcomes may be affected[34]. For the best model identified in this effort, an FPR of 22.5 % was achieved. While the GFF ANN had a smaller FPR of 18.3 %, there was a decrease in diagnostic accuracy. A final limitation, especially as it concerns future research, is the fact that significant multi-system organ failure and the LFRI may have been responding to severe liver dysfunction or other organ fail-

ure that was undiagnosed in the patient.

Conclusion

This effort resulted in the development of machine learning-based models and a novel intuitive LFRI that hold significant promise for enhancing the diagnosis and prediction of liver dysfunction/failure in a critical care patient population. To establish potential clinical utility, the developed models were extensively validated using an open-access critical care database. The developed models and LFRI could be integrated into future technology or software applications, including clinical decision support systems to assist health care providers in early identification liver dysfunction/failure in the critical care setting that can contribute to substantial improvements in healthcare delivery, patient safety, and outcomes.

Conflict of Interest

All authors declare that they have no conflict of interests to report. The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

Data used throughout this effort was obtained from two open access critical care databases MIMIC-III and eICU Collaborative Research databases as referenced in the manuscript.

Funding was received to complete the referenced study and research.

References

1. Said HM. Physiology of the Gastrointestinal Tract. 6th Edition. Academic Press, Elsevier, Cambridge, MA, 2018.
2. de Ataide EC, Reges Perales S, de Oliveira Peres MA, Bastos Eloy da Costa L, Quarella F, Valerini FG, et al. Acute Liver Failure Induced by Carthamus tinctorius Oil: Case Reports and Literature Review. Transplant Proc. 2018; 50: 476-477.
3. Stravitz RT, Kramer DJ. Management of acute liver failure. Nat Rev Gastroenterol Hepatol. 2009; 6: 542-553.
4. Rajaram P, Subramanian R. Management of Acute Liver Failure in the Intensive Care Unit Setting. Clin Liver Dis. 2018; 22: 403-408.
5. Casey LC, Fontana RJ, Aday A, Nelson DB, Rule JA, Gottfried M, et al. Acute Liver Failure (ALF) in Pregnancy: How Much Is Pregnancy Related? Hepatology. 2020; 72: 1366-1377.
6. Singh T, Gupta N, Alkhouri N, Carey WD, Hanouneh IA. A guide to managing acute liver failure. Cleve Clin J Med. 2016; 83: 453-462.
7. Mawatari S, Harada Y, Iwai M, Kwo PY, Ido A. Acute Liver Failure. Diagnosis of Liver Disease. McGraw Hill, New York, 2019, 45-50.
8. Nanchal R, Subramanian R, Karvellas CJ, Hollenberg SM, Peppard WJ, Singbartl K, et al. Guidelines for the Management of Adult Acute and Acute-on-Chronic Liver Failure in the ICU: Cardiovascular, Endocrine, Hematologic, Pulmonary, and Renal Considerations. Crit Care Med. 2020; 48: e173-e191.
9. Shingina A, Mukhtar N, Wakim-Fleming J, Alqahtani S, Wong RJ, Limketkai BN, et al. Acute Liver Failure Guidelines. Am J Gastroenterol. 2023; 118: 1128-1153.
10. Kwo PY, Cohen SM, Lim JK. ACG Clinical Guideline: Evaluation of Abnormal Liver Chemistries. Am J Gastroenterol. 2017; 112: 18-35.
11. El Hadi H, Di Vincenzo A, Vettor R, Rossato M. Relationship between Heart Disease and Liver Disease: A Two-Way Street. Cells. 2020; 9: 567.
12. Manikat R, Nguyen MH. Nonalcoholic fatty liver disease and non-liver comorbidities. Clin Mol Hepatol. 2023; 29: s86-s102.
13. Horsfall LJ, Clarke CS, Nazareth I, Ambler G. The value of blood-based measures of liver function and urate in lung cancer risk prediction: A cohort study and health economic analysis. Cancer Epidemiol. 2023; 84: 102354.
14. Chen EQ, Zeng F, Zhou LY, Tang H. Early warning and clinical outcome prediction of acute-on-chronic hepatitis B liver failure. World J Gastroenterol. 2015; 21: 11964-11973.
15. Abdar M, Yen NY, Hung JCS. Improving the Diagnosis of Liver Disease Using Multilayer Perceptron Neural Network and Boosted Decision Trees. J Med Biol Eng. 2018; 38: 953-965.
16. Hydes TJ, Meredith P, Schmidt PE, Smith GB, Prytherch DR, Aspinall RJ. National Early Warning Score Accurately Discriminates the Risk of Serious Adverse Events in Patients With Liver Disease. Clin Gastroenterol Hepatol. 2018; 16: 1657-1666.e10.
17. Zare A, Zare MA, Zarei N, Yaghoobi R, Zare MA, Salehi S, et al. A Neural Network Approach to Predict Acute Allograft Rejection in Liver Transplant Recipients Using Routine Laboratory Data. Hepat Mon. 2017; 17: e55092.
18. Lee H, Yoon S, Oh SY, Shin J, Kim J, Jung CW, et al. Comparison of APACHE IV with APACHE II, SAPS 3, MELD, MELD-Na, and CTP scores in predicting mortality after liver transplantation. Sci Rep. 2017; 7: 10884.
19. Cholongitas EB, Betrossian A, Leandro G, Shaw S, Patch D, Burroughs AK. King's criteria, APACHE II, and SOFA scores in acute liver failure. Hepatology. 2006; 43: 881; author reply 882.
20. Katoonizadeh A, Laleman W, Verslype C, Wilmer A, Maleux G, Roskams T, et al. Early features of acute-on-chronic alcoholic liver failure: a prospective cohort study. Gut. 2010; 59: 1561-1569.
21. de Liguori Carino N, O'Reilly DA, Dajani K, Ghaneh P, Poston GJ, Wu AV. Perioperative use of the LiMON method of indocyanine green elimination measurement for the prediction and early detection of post-hepatectomy liver failure. Eur J Surg Oncol. 2009; 35: 957-962.
22. Costa E Silva PP, Codes L, Rios FF, Esteve CP, Valverde Filho MT, Lima DOC, et al. Comparison of General and Liver-Specific Prognostic Scores in Their Ability to Predict Mortality in Cirrhotic Patients Admitted to the Intensive Care Unit. Can J Gastroenterol Hepatol. 2021; 2021: 9953106.
23. Thüring J, Rippel O, Haarburger C, Merhof D, Schad P, Bruners P, et al. Multiphase CT-based prediction of Child-Pugh classification: a machine learning approach. Eur Radiol Exp. 2020; 4: 20.
24. Yip TC, Ma AJ, Wong VW, Tse YK, Chan HL, Yuen PC, et al. Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (NAFLD) in the general population. Aliment Pharmacol Ther. 2017; 46: 447-456.
25. Anagnostou T, Remzi M, Lykourinas M, Djavan B. Artificial neural networks for decision-making in urologic oncology. Eur Urol. 2003; 43: 596-603.
26. Pournik O, Dorri S, Zabolinezhad H, Alavian SM, Eslami S. A diagnostic model for cirrhosis in patients with non-alcoholic fatty liver disease: an artificial neural network approach. Med J Islam Repub Iran. 2014; 28: 116.
27. Perez Ruiz de Garibay A, Kortgen A, Leonhardt J, Zipprich A, Bauer M. Critical care hepatology: definitions, incidence, prognosis and role of liver failure in critically ill patients. Crit Care. 2022; 26: 289.
28. Vasques F, Cavazza A, Bernal W. Acute liver failure. Curr Opin Crit Care. 2022; 28: 198-207.

29. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016; 3: 160035.
30. Nasa P, Jain R, Juneja D. Delphi methodology in healthcare research: How to decide its appropriateness. World J Methodol. 2021; 11: 116-129.
31. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018; 5: 180178.
32. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. Crit Care Med. 2018; 46: 547-553.
33. Bin Rafiq R, Modave F, Guha S, Albert MV. Validation Methods to Promote Real-world Applicability of Machine Learning in Medicine. 3rd International Conference on Digital Medicine and Image Processing (DMIP' 2020). Available at: https://dl.acm.org/doi/fullHtml/10.1145/3441369.3441372, date accessed: 20/12/2023.
34. Ruskin KJ, Hueske-Kraus D. Alarm fatigue: impacts on patient safety. Curr Opin Anaesthesiol. 2015; 28: 685-690.